

Ricco Rakotomalala

Comparaison de populations

Tests non paramétriques

Version 1.0

Université Lumière Lyon 2

Avant-propos

Ce support est dédié aux tests non paramétriques de comparaison de populations. Il vient en contre-point du fascicule consacré aux tests paramétriques (http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf). Ces tests répondent à des problématiques identiques mais s'appuient sur des mécanismes différents. Il ne faudra donc pas s'étonner de voir des parties communes entre les 2 documents lorsque nous présentons les problèmes à analyser. Autant que possible, nous ferons référence à son équivalent paramétrique lorsque nous présenterons une technique non paramétrique.

La grande majorité des techniques présentées dans ce support ont été implémentées dans le logiciel TANAGRA, un outil *open source* accessible en ligne <http://eric.univ-lyon2.fr/~ricco/tanagra/>. Des tutoriels peuvent être consultés sur un site dédié <http://tutoriels-data-mining.blogspot.com/>. Ils sont regroupés par thèmes, une recherche par mot-clés est possible. D'autres tutoriels concernant d'autres techniques, les tests paramétriques notamment, sont également disponibles.

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 (on utilise également l'appellation *test d'homogénéité*) 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 teneurs en sel de chaque type de hamburger (cf. *comparaison de caractéristiques de localisation, échantillons indépendants*).
- Évaluer la réduction de la variabilité des pièces produites par une machine en introduisant de nouvelles normes de qualité (cf. *comparaison de caractéristiques de dispersion*).
- 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.

Non paramétrique. On parle de tests non paramétriques lorsque l'on ne fait aucune hypothèse sur la distribution des variables. On parle également de tests **distribution free** c.-à-d. la qualité des résultats ne dépend pas, *a priori*, de la distribution sous-jacente des données. Il faut simplement que sous l'hypothèse nulle, les fonctions de répartitions, conditionnellement aux sous populations, soient identiques. L'hypothèse alternative générique est leur différence. On peut affiner les résultats en précisant la forme de cette différence (caractéristiques de localisation ou d'échelle). Première conséquence de cette liberté supplémentaire, les tests non paramétriques s'appliquent aux variables d'intérêt (variable à analyser) quantitatives, mais aussi aux variables ordinales, en fait tout type de données qui permette de classer les observations (ex. notes attribuées, classement par des juges, degré de préférence, ordre de grandeur, etc.). Nous préciserons plus loin les principaux avantages et inconvénients des tests non paramétriques.

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. Trouver des sites web via un moteur de recherche est chose aisée. Après, vient un long travail de vérification, de recoupement, d'implémentation (reproduire les calculs "à la main" sous Excel est la meilleure manière de comprendre une technique à mon sens), toujours fastidieux, mais indispensable si on veut s'assurer de l'exactitude des formulations retranscrites. Le plus souvent, ces sources seront directement indiquées dans le texte, en note de bas de page. Dans certains cas, quand il s'agit de sites de cours complets, les références seront insérées dans la bibliographie.

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. L'ouvrage de Siegel et Castellan [13]. Cet ouvrage a été une véritable découverte pour moi, tant sur les tests non paramétriques que sur la manière de présenter un sujet. J'ai essayé de m'en inspirer lorsque j'ai commencé à écrire mes propres supports. Les auteurs utilisent, de manière systématique, la trame suivante : (a) de quoi s'agit-il ; (b) pourquoi et dans quel cadre peut-on mettre en oeuvre la technique ; (c) quel est le détail des formules ; (d) dérouler le processus sur un petit exemple, en insistant sur les étapes clés ; (e) comment lire les résultats ; (f) que faut-il en penser finalement, dans quel contexte la

méthode est la plus (moins) efficace par rapport aux techniques existantes. La lecture de cet ouvrage m'a énormément aidé pour avoir les idées claires concernant les tests non paramétriques¹.

2. L'ouvrage de Capéraà et Van Cutsem [3]. Cet ouvrage a été mon premier contact avec les tests non paramétriques. Il a failli être le dernier. Rarement j'ai vu un texte aussi touffu et aride. Je ne sais pas si ça tient à l'écriture serrée, aux formules difficiles à déchiffrer, à l'absence d'exemples didactiques, mais l'étudiant que j'étais a très vite baissé les bras. Quelques années plus tard, avec un peu plus de recul sur les techniques (un peu plus d'expérience peut être aussi), je me suis replongé dans cet ouvrage... et je me suis rendu compte de son exceptionnelle qualité scientifique. C'est une vraie mine d'or. Manifestement, de mon point de vue, les auteurs s'adressent aux spécialistes et non aux praticiens. Une fois cette idée admise, on se rend compte de la richesse des informations prodiguées : tous les théorèmes sont donnés et démontrés. Une véritable monographie pour chercheurs. C'est simple : à chaque fois que j'ai eu un doute sur une formulation lors de l'élaboration de ce support, j'ai utilisé en priorité cette référence.
3. Le site web <http://hdelboy.club.fr/Nonparam.htm>. Ce site est très complet. Il présente les principaux tests non paramétriques. Et surtout il montre leur mise en oeuvre sous le tableur Excel en détaillant les calculs, exactement ce que je m'évertue à faire dans mes propres supports. Des macros complémentaires facilitent la prise en main des tests. Autre avantage indéniable de ce site, toutes les sources sont fournies et un très grand nombre de liens vers d'autres références en ligne sont disponibles. On remarquera en particulier les liens vers les tables 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.

¹ Une adaptation et traduction de cet ouvrage (entres autres) est accessible en ligne : R. Rasmousse, M. Le Berre, L. Le Guelte, *Introduction aux statistiques*, <http://www.cons-dev.org/elearning/stat/index.html> et <http://www.cons-dev.org/elearning/ando/index.html>

Table des matières

Partie I Tests pour échantillons indépendants

1	Tests génériques de comparaison pour $K = 2$ échantillons	11
1.1	Test de Kolmogorov-Smirnov	14
1.1.1	Principe, statistique du test et région critique	14
1.1.2	Approximations pour les grands échantillons	16
1.1.3	Le cas particulier $n_1 = n_2 = m$	21
1.1.4	Le test de Kuiper	22
1.2	Le test de Cramer - von Mises	23
1.2.1	Principe, statistique de test et région critique	23
1.2.2	Traitement des ex-aequo	27
1.3	Lecture des sorties des logiciels	27
1.3.1	Comparer les salaires féminins	27
1.3.2	Comparer les salaires masculins	30
2	Tests de rang dans un modèle de localisation pour $K = 2$ échantillons	33
2.1	Test de Wilcoxon - Mann - Whitney	34
2.1.1	Rangs, somme des rangs et traitement des ex-aequos	34
2.1.2	Statistiques de rang linéaires	37
2.1.3	Statistique de test et région critique	38
2.1.4	Loi asymptotique - L'approximation normale	39
2.1.5	Correction pour les ex-aequo	41
2.1.6	La variante de Wilcoxon	44
2.1.7	Une autre vision du test de Mann et Whitney	45

2.1.8	Intérêt du test de Wilcoxon-Mann-Whitney	46
2.1.9	Sorties des logiciels	46
2.2	Test de Fisher - Yates - Terry - Hoeffding (FYTH)	49
2.2.1	Principe, statistique de test	49
2.2.2	Approximation normale pour les grands effectifs	50
2.3	Test de Van der Waerden	52
2.4	Test de rang robuste	56
2.4.1	Le problème de Behrens-Fisher	56
2.4.2	Test de rang robuste de Fligner-Policello	56
2.5	Test de la médiane	60
2.5.1	Principe, statistique de test et région critique (Approche A)	60
2.5.2	Approche par les statistiques de rang linéaires (Approche B)	61
2.5.3	Intérêt du test de la médiane	62
2.5.4	Un exemple et deux approches	63
3	Tests de rang dans un modèle de localisation pour $K \geq 2$ populations	67
3.1	Test de Kruskal-Wallis	67
3.1.1	Principe, statistique de test et région critique	67
3.1.2	Distribution asymptotique	70
3.1.3	Traitement des ex-aequo	71
3.1.4	Sorties des logiciels	74
3.2	Détermination de la source des écarts	75
3.2.1	Comparaisons multiples	76
3.2.2	Comparaisons à une référence	77
3.3	Autres tests pour $K \geq 2$	79
3.3.1	Autant de tests que de fonctions score	79
3.3.2	Le test de la médiane généralisée	80
3.4	Tests pour les alternatives ordonnées (modèle de localisation)	83
3.4.1	Position du problème	83
3.4.2	Test de Jonckheere-Terpstra pour échantillons indépendants	84
3.4.3	Test de Page pour échantillons indépendants	88

4	Tests de rang dans un modèle d'échelle	91
4.1	Test de Mood	92
4.1.1	Principe, statistique de test	92
4.1.2	Approximation par la loi normale	93
4.1.3	Correction pour les ex-aequo	94
4.1.4	Lorsque les paramètres de localisation sont différents	95
4.2	Test de Klotz	95
4.2.1	Principe du test de Klotz	95
4.2.2	Quelques considérations sur les performances	97
4.2.3	Sorties des logiciels	97
4.3	Test robuste de Moses	99
4.3.1	Construction du test	99
4.3.2	Quelques commentaires	102
4.4	Généralisation à $K \geq 2$ groupes	102
5	Retour sur les statistiques de rang linéaires	105
5.1	Statistiques de rang linéaires	105
5.2	Écriture des statistiques de test pour ($K = 2$)	106
5.3	Écriture des statistiques de test pour ($K \geq 2$)	107
5.4	Les principales fonctions scores	107
5.4.1	Score de Wilcoxon	108
5.4.2	Score de la médiane	108
5.4.3	Score de Van der Waerden	109
5.4.4	Score de Mood	109
5.4.5	Score de Klotz	109
5.4.6	Score de Savage	110
5.4.7	Score de Siegel-Tukey	110
5.4.8	Score de Ansari-Bradley	110
5.4.9	Traitement des ex-aequo	111
5.4.10	Comparaison des fonctions score sur données simulées	111
5.5	Quelques exemples (repris des chapitres précédents)	114
5.5.1	Test de Wilcoxon-Mann-Whitney	115
5.5.2	Test de Van der Waerden	116
5.5.3	Test de Kruskal-Wallis	118

Partie II Tests pour échantillons appariés

6	Tests pour ($K = 2$) échantillons liés	123
6.1	Principe de l'appariement	123
6.2	Test des signes	124
6.2.1	Test d'hypothèses, statistique de test et région critique	124
6.2.2	L'approximation normale pour les grands effectifs	127
6.2.3	Sorties des logiciels	128
6.3	Test des rangs signés de Wilcoxon	130
6.3.1	Test d'hypothèses, statistique de test et région critique	130
6.3.2	Approximation normale pour les grands effectifs	134
6.3.3	Sorties des logiciels	137
7	Tests pour ($K \geq 2$) échantillons liés	141
7.1	Appariement pour $K \geq 2$ échantillons - Les blocs aléatoires complets	141
7.2	ANOVA de Friedman	142
7.2.1	Principe, statistique de test et région critique	142
7.2.2	Approximations pour les grands échantillons	145
7.2.3	Traitement des ex-aequo	146
7.2.4	Sorties des logiciels	148
7.3	Détermination de la source des écarts	149
7.3.1	Comparaisons multiples	149
7.3.2	Comparaisons à une référence	150
7.4	Tests pour les alternatives ordonnées	152
7.4.1	Test de Page pour échantillons liés	153
7.4.2	Test de Jonckheere pour échantillons liés	155

8 Tests pour les variables binaires	159
8.1 Test de McNemar pour la comparaison de $K = 2$ populations	159
8.1.1 Principe, statistique de test et région critique	159
8.1.2 Une approche non symétrique du test de McNemar	162
8.2 Le test de Stuart-Maxwell : généralisation de McNemar à L modalités	164
8.2.1 Principe, statistique de test et région critique	164
8.2.2 La variante de Bhapkar	167
8.3 Test Q de Cochran pour $K \geq 2$ populations	169
8.3.1 Principe, statistique de test et région critique	169
8.3.2 Sorties des logiciels	171
8.3.3 Détecter la source des écarts	172
A Gestion des versions	175
B Les tests non paramétriques avec le logiciel TANAGRA	177
C Les tests non paramétriques avec d'autres logiciels libres	179
D Tables statistiques	181
D.1 Test de Kolmogorov-Smirnov	182
D.2 Test de Cramer - von Mises	183
D.3 Table de Mann et Whitney	184
D.4 Table de Kruskal et Wallis	185
D.5 Valeurs critiques du test de rangs signés de Wilcoxon pour échantillons appariés	186
D.6 Table des valeurs critiques de la statistique de Friedman pour la comparaison de K échantillons liés	187
Littérature	189

Pourquoi les tests non paramétriques ?

Au lieu d'entrer dans le débat stérile pour ou contre les tests non paramétriques en les opposant à leurs homologues paramétriques fondés sur la distribution normale des données, essayons plutôt de caractériser les situations où il est plus (ou moins) avantageux de les utiliser.

Pas d'hypothèses contraignantes sur les distributions

Le principal atout des tests non paramétriques est d'être *distribution free*. Il n'est pas nécessaire de faire des hypothèses sur la forme des distributions, il n'est pas nécessaire non plus d'estimer les paramètres associés (ex. la moyenne si la répartition était gaussienne, etc.). Le champ d'application des techniques est par conséquent plus large. Vérifier a priori les conditions de validité des tests n'est pas un préalable indispensable.

Un premier point de vue vient relativiser cet avantage. Les tests paramétriques sont assez robustes, même si on s'écarte des conditions d'utilisation. L'analyse de variance par exemple conserve des qualités raisonnables même si la normalité sous jacente n'est pas respectée, pour peu que les effectifs soient assez élevés ; et même si les variances conditionnelles sont sensiblement différentes, pour peu que les effectifs soient équilibrés. De plus, il est toujours possible de transformer les variables de manière à se rapprocher de la distribution normale². Et lorsque les conditions d'application sont respectées, les tests paramétriques sont les plus puissants.

A contrario, un second point de vue vient confirmer l'avantage des tests non paramétriques. En effet, même lorsque les conditions d'applications des tests paramétriques sont réunies, l'avantage de ces derniers par rapport aux tests non paramétriques n'est pas transcendant. Pour reprendre une configuration maintes fois citée dans la littérature, lorsque la variable d'intérêt est gaussienne, l'efficacité relative asymptotique du test de Wilcoxon-Mann-Whitney (le test non paramétrique certainement le plus populaire) par rapport au test de Student est de $\frac{3}{\pi} \approx 95\%$ c.-à-d. lorsque l'hypothèse alternative est vraie, les moyennes sont différentes, s'il faut 95 observations pour aboutir à cette conclusion avec le test de Student, il en faudra 100 avec le test non paramétrique. Dans tous les autres types de loi, ce dernier est meilleur (voir [3], page 154 ; [13], page 137 ; <http://www.jerrydallal.com/LHSP/npar.htm>). Quant au test de Van der Waerden, certes moins connu, il est simplement aussi bon que le test de Student lorsque la variable d'intérêt est gaussienne (voir [3], page 154 ; [1], page 319).

De fait, dans une phase d'investigation où l'on cherche à distinguer les informations que recèlent les données, s'affranchir d'hypothèses contraignantes, la forme des distributions en est une, est certainement un atout supplémentaire.

Tests adaptés aux variables ordinales

Les tests non paramétriques sont naturellement adaptés aux données ordinales. Bien souvent les données disponibles sont des ordres de grandeurs. Il est difficile de donner une interprétation claire à

² Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf concernant les techniques de transformation, notamment les formules de Box-Cox

la valeur. Dans ce cas, le plus important est la position relative des observations, leur ordonnancement. Prenons les notes obtenues par 3 étudiants à l'examen : le premier a obtenu 15/20, le second 13/20 et le troisième 12/20. Peut-on vraiment dire que l'écart entre le premier et le second est 2 fois plus important que l'écart entre le second et le troisième? Et pourtant c'est ce que nous ferons dès que nous voudrions réaliser des calculs statistiques basées sur les distances. En transformant les données en rangs, les tests non paramétriques s'affranchissent de ce type de problème.

Notons quand même que nous perdons de l'information en transformant les données en rangs. Il faut en être conscient. Si les notes des étudiants avaient été respectivement 15/20, 6/20 et 3/20, dès que l'on passe aux rangs, les deux configurations sont traitées de manière identique alors qu'elles sont manifestement différentes.

Enfin, lorsque les données s'expriment comme des degrés de préférence (ex. très insatisfait, insatisfait, indifférent, satisfait, très satisfait), fondamentalement ordinales avec peu de valeurs possibles, les tests non paramétriques sont les seuls applicables. C'est également le cas lorsque les données sont catégorielles (voir le support concernant les "Mesures d'association entre variables qualitatives" - http://eric.univ-lyon2.fr/~ricco/cours/cours/Dependance_Variables_Qualitatives.pdf).

Robustesse par rapport aux points aberrants

La transformation des données en rangs introduit une propriété très appréciable lorsque nous traitons des problèmes réels : les statistiques sont bien moins sensibles aux points aberrants. En effet, la présence d'un point atypique fausse très souvent la moyenne, qui joue un rôle central dans les tests paramétriques. Si le point atypique correspond à une très grande valeur, s'écartant fortement des autres, la moyenne est "tirée" vers le haut, biaisant tous les calculs subséquents. Avec les rangs, on utilise uniquement l'information "le point correspond à la valeur la plus élevée", le rôle néfaste de l'observation atypique est amoindri.

Tests adaptés aux petits échantillons

Lorsque les effectifs sont faibles, les tests paramétriques, à moins vraiment que l'hypothèse de normalité ne soit établie, ne sont pas opérantes, à la différence des tests non paramétriques. Ces derniers sont alors sans concurrence. Des tables statistiques spécifiques sont disponibles pour réaliser des tests exacts.

Additionnée à cette première qualité, la convergence vers les lois asymptotiques est très rapide. Dans la pratique, dès que les effectifs atteignent un niveau modéré (de l'ordre de 20 à 30 observations, cela dépend du test), les approximations sont déjà pleinement efficace. Pour le test de Wilcoxon-Mann-Whitney par exemple, il suffit que $n_1 > 10$ (ou $n_2 > 10$) pour que l'approximation normale soit valable (voir [13], page 132; on propose $n_1 + n_2 > 20$ avec $n_1 > 3$ et $n_2 > 3$ dans [2], page 19; ou même $n_1 > 8$ et $n_2 > 8$ dans [1], page 317).

Tests plus souples

Statistiques non paramétriques ne veut pas dire absence totale de l'étude des paramètres (je préfère le terme "caractéristique") des distributions lors de la mise en place des tests. Cela indique surtout que le test ne repose pas sur les paramètres des présumées distributions des données, *définis à l'avance*, et que l'on doit estimer sur les échantillons (la moyenne et la variance si on se réfère à la loi normale).

De fait, les hypothèses que l'on peut confronter sont plus riches : nous pouvons tester la différence brute entre les fonctions de répartition, tout type d'écart correspond au rejet de l'hypothèse nulle (test de Kolmogorov-Smirnov) ; comparer les caractéristiques de localisation ou le décalage entre les fonctions de répartition (test des médianes, test de Wilcoxon-Mann-Whitney) ; comparer les caractéristiques de dispersion ou d'échelle des observations (test de Klotz).

L'hypothèse nulle est donc toujours l'égalité des fonction de répartition. Mais selon le résultat que nous souhaitons mettre en évidence, nous définissons le type de caractéristique que nous souhaitons étudier (caractéristiques de localisation ou d'échelle). Le rejet de l'hypothèse nulle est le fruit de la différenciation selon cette caractéristique.

Notations

Les données proviennent de K échantillons Ω_k ($k = 1, \dots, K$). La variable X est notée en majuscule, la valeur pour l'observation n^oi 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.

Lors de la transformation des données en rangs, nous noterons r_i le rang de l'observation n^oi .

L'effectif global est n , les sous-échantillons comportent n_k observations, avec $n = n_1 + \dots + n_K$.

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^o1 sera μ_1 , etc.). La valeur de l'individu n^oi dans l'échantillon Ω_k sera notée x_{ik} .

Concernant les rangs, ils sont toujours calculés sur la totalité de l'échantillon. Pour l'individu n^oi de l'échantillon Ω_k , le rang r_{ik} correspond à la position de la valeur x_{ik} dans l'ensemble des observations Ω .

Données

Contrairement au support dédié aux tests paramétriques de comparaison de populations³, nous illustrerons les techniques à l'aide de données spécifiques selon les chapitres et les sections. Le passage des

³ http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf

données au rangs et le traitement des ex-aequo sont des mécanismes importants pour les tests non paramétriques, ils méritent que l'on s'y attarde un peu plus. Le discours sera plus facile à faire passer si les données sont faciles à appréhender, avec des effectifs faibles et des situations bien délimitées.

Néanmoins, nous utiliserons également le fichier de demandeurs de crédits mis à contribution dans le document consacré aux tests paramétriques. Il servira à illustrer la mise en oeuvre des tests sous différents logiciels. Nous en profiterons également pour comparer les résultats à ceux des tests paramétriques homologues. Rappelons que nous avons vérifié la normalité des variables, toutes sont compatibles avec une distribution gaussienne. Cela laisse à penser que les tests paramétriques devraient être les plus puissants dans toutes les configurations que nous aurons à étudier.

Pour rappel, 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) :

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)

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

Numéro	Sal.Homme	Sal.Femme	Rev.Tete	Age	Acceptation	Garantie.Supp	Emploi
1	7.92	7.72	7.42	3.69	oui	hypothèque	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	hypothèque	cdd
6	6.89	6.51	6.72	4.16	non	hypothèque	cdd
7	7.29	6.93	6.43	3.37	oui	hypothèque	cdd
8	7.53	7.51	7.52	3.99	oui	hypothèque	cdd
9	7.48	7.25	6.46	3.47	oui	non	cdi
10	7.27	6.60	6.59	3.30	oui	hypothèque	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	hypothèque	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	hypothèque	cdi
18	6.83	7.06	6.03	3.74	oui	hypothèque	cdi
19	6.98	7.29	6.74	3.83	non	hypothèque	cdd
20	7.80	7.38	7.61	3.97	oui	hypothèque	cdi
21	7.67	7.69	7.27	3.81	oui	hypothèque	cdi
22	7.28	7.05	7.17	3.30	oui	caution	cdi
23	7.17	6.86	6.62	3.40	non	hypothèque	cdd
24	7.42	7.25	6.42	3.40	non	non	cdd
25	7.83	7.77	7.40	3.76	oui	hypothèque	cdi
26	7.33	7.14	7.24	3.18	oui	hypothèque	cdi
27	6.02	6.03	5.11	3.26	oui	hypothèque	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	hypothèque	cdi
32	8.03	7.94	7.29	3.78	oui	non	cdi
33	8.46	8.12	8.30	3.69	oui	hypothèque	cdi
34	6.64	7.12	6.22	3.50	oui	hypothèque	cdi
35	7.92	7.92	6.82	3.66	oui	non	cdi
36	7.14	7.20	6.26	3.78	non	hypothèque	cdd
37	7.13	6.85	6.08	3.85	non	caution	cdd
38	7.43	7.20	7.32	4.11	oui	hypothèque	cdi
39	8.78	8.58	8.69	3.78	oui	non	cdi
40	8.28	7.85	7.68	3.74	oui	hypothèque	cdi
41	6.31	6.57	5.75	3.66	non	hypothèque	cdd
42	7.48	6.97	7.26	3.74	non	hypothèque	cdd
43	7.48	6.96	6.85	3.37	non	hypothèque	cdi
44	7.69	7.11	7.44	4.16	non	hypothèque	cdi
45	7.44	7.16	6.91	3.78	non	non	cdi
46	7.47	7.24	6.45	3.66	oui	hypothèque	cdi
47	8.17	8.29	8.23	3.95	oui	non	cdi
48	7.40	7.29	7.35	3.09	non	hypothèque	cdi
49	7.26	6.81	7.06	4.16	non	non	cdi
50	7.50	7.16	7.35	4.08	non	hypothèque	cdi

Fig. 0.1. Fichier des demandeurs de crédits

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 $n_1 = n_2 = \dots = n_K$ 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 Ω_j ($j \neq k$). 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.*).

Tests génériques de comparaison pour $K = 2$ échantillons

Lors de la comparaison de $K = 2$ échantillons, nous cherchons à savoir si les observations proviennent de la même population au regard de la variable d'intérêt ; ou, de manière équivalente, la distribution de la variable d'intérêt est la même dans les 2 sous échantillons.

Les tests génériques d'homogénéité visent à **détecter toute forme différenciation entre les 2 distributions empiriques**. Cela peut être un décalage entre les distributions, une différence de tendance centrale (paramètre de localisation), une dispersion différente (paramètre d'échelle), des queues de distribution plus ou moins lourdes, une asymétrie ou un aplatissement différents, etc.

Il y a une contrepartie à cela. Étant très généraux, ils sont peu puissants et n'expliquent pas la nature de la différenciation.

Le test d'hypothèse s'écrit :

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

où $F_k(x)$ est la fonction de répartition de la variable d'intérêt X dans la sous population associée à l'échantillon Ω_k .

Remarque 1 (Test unilatéral à gauche ou à droite). Il est possible de définir un test unilatéral. Pour traduire l'idée "les valeurs prises par X ont elles tendance à être plus petites dans la première sous population ?", nous définirons le test de la manière suivante :

$$H_0 : F_1(X) = F_2(X)$$

$$H_1 : F_1(X) > F_2(X)$$

Le sens de l'inégalité dans l'hypothèse alternative peut paraître étrange si l'on se réfère au test que l'on veut mettre en place. Nous préciserons cela sur un exemple concret dans ce qui suit.

Quelques illustrations

Nous utilisons des données simulées pour illustrer différentes configurations.

Écart selon le paramètre de localisation. Dans la figure 1.1, nous avons généré 2 échantillons de taille $n_1 = n_2 = 100$. Nous avons tracé les fonctions de densité empiriques respectives. Le premier échantillon Ω_1 (trait en pointillé) correspond à une loi normale $\mathcal{N}(0;1)$, le second Ω_2 (trait continu) à une $\mathcal{N}(1;1)$. Bien évidemment, les valeurs dans le second échantillon sont stochastiquement plus élevées que celles du premier.

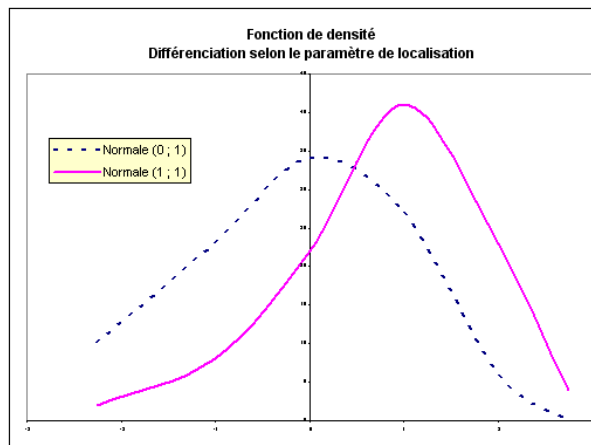


Fig. 1.1. Fonctions de densité conditionnelles - Différence de paramètre de localisation

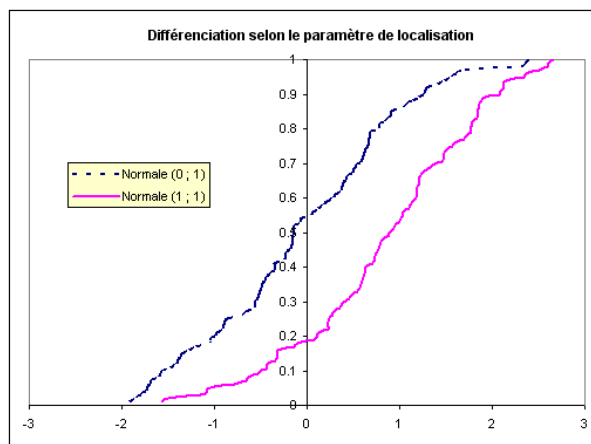


Fig. 1.2. Fonctions de répartition conditionnelles - Différence de paramètre de localisation

Concernant les fonctions de répartition, on constate que $F_1(X)$ est systématiquement au dessus de $F_2(X)$. Pour bien comprendre ce mécanisme, prenons comme référence les quantiles de la variable d'intérêt : ils sont systématiquement plus petits dans le premier échantillon. Par exemple, pour le quantile

d'ordre 0.1, $Q_1(0.1) = -1.55$ et $Q_2(0.1) = -0.51$; pour le quantile d'ordre 0.2, $Q_1(0.2) = -0.99 < Q_2(0.2) = 0.11$; ...; la médiane, $Q_1(0.5) = -0.14 < Q_2(0.5) = 0.89$; etc.

Les 2 fonctions de distributions sont décalées au regard du paramètre de localisation ou de caractéristique de tendance centrale : X a tendance à prendre des valeurs plus élevées dans le deuxième sous échantillon, on a bien $F_1(X) > F_2(X)$.

Écart selon le paramètre d'échelle. Voyons un second exemple où les distributions se différencient selon leur caractéristique d'échelle. Le premier échantillon correspond à une $\mathcal{N}(0;1)$, les valeurs observées varient entre $[-1.88; 3.12]$; le second à une $\mathcal{N}(0;0.2)$, les valeurs observées sont comprises entre $[-0.78; 0.51]$ (Figure 1.3).

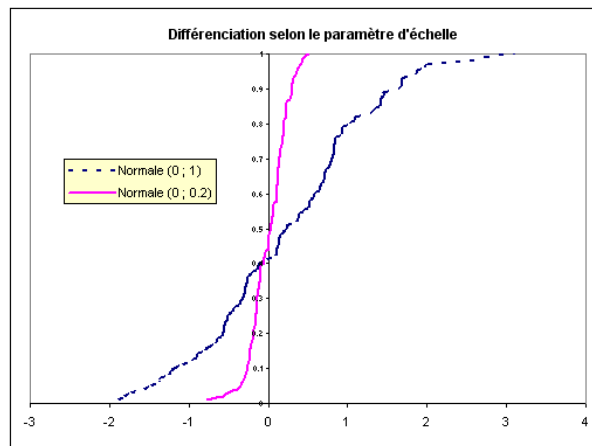


Fig. 1.3. Fonctions de répartition conditionnelles - Différence de paramètre d'échelle

Les deux fonctions de répartition empirique se croisent. La seconde est plus resserrée autour de sa caractéristique de tendance centrale (moyenne, médiane). L'écart entre les fonctions de répartition est d'abord positif avant d'être négatif. Concrètement, concernant les quantiles

$$Q_1(\alpha) < Q_2(\alpha) \text{ , pour } \alpha < 0.5$$

$$Q_1(\alpha) > Q_2(\alpha) \text{ , pour } \alpha > 0.5$$

Compte tenu des caractéristiques des valeurs générées dans la simulation, les médianes observées sont très proches (les médianes théoriques sont identiques).

1.1 Test de Kolmogorov-Smirnov

1.1.1 Principe, statistique du test et région critique

Statistique de test

Le test de Kolmogorov Smirnov vise à détecter toute forme de différenciation entre les distributions. Il repose sur l'écart maximum entre les fonctions de répartition empiriques. Pour un test bilatéral, la statistique du test s'écrit

$$D = \max_x |F_1(x) - F_2(x)| \quad (1.1)$$

Pour les tests unilatéraux, nous utiliserons selon le cas

$$D^+ = \max_x [F_1(x) - F_2(x)] \quad (1.2)$$

$$D^- = \max_x [F_2(x) - F_1(x)] \quad (1.3)$$

On constate alors que

$$D = \max(D^+, D^-)$$

Il s'agit de mesurer **l'écart vertical** entre les fonctions de répartition. Si nous reprenons notre exemple des données simulées avec des paramètres de localisation différents (Figure 1.4), la valeur de D correspondrait¹ à l'écart entre les ordonnées au point $x \approx 0.1$, soit $D \approx |0.56 - 0.19| = 0.37$.

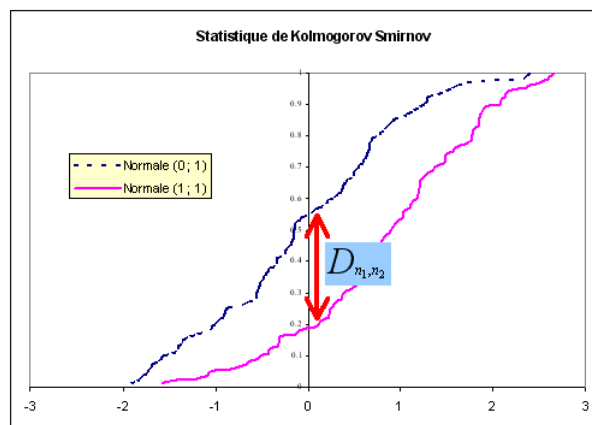


Fig. 1.4. Détermination de la statistique de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est convergent pour toute hypothèse alternative de la forme $F_1(X) \neq F_2(X)$ c.-à-d. la probabilité de rejeter H_0 tend vers 1 lorsque $n_1, n_2 \rightarrow \infty$. Néanmoins, comme tous les tests omnibus c.-à-d. censés détecter toute forme de différenciation, il est peu puissant, avec un risque de deuxième espèce élevé. Il conclut un peu trop souvent à la compatibilité des données avec l'hypothèse nulle alors que l'hypothèse alternative est vraie.

¹ Tout cela de manière visuelle, très approximative, nous décrirons les stratégies de calcul dans la section suivante

Région critique

Pour le test bilatéral, nous rejetons l'hypothèse nulle lorsque l'écart maximum mesuré est anormalement élevé. La région critique du test au risque α s'écrit

$$R.C. : D \geq k_\alpha(n_1, n_2)$$

Où $k_\alpha(n_1, n_2)$ est lue dans la table des valeurs critiques de Kolmogorov-Smirnov.

Remarque 2 (Tests unilatéraux). Le mécanisme est le même pour les tests unilatéraux. Les valeurs élevées sont suspectes, les valeurs critiques seront différentes simplement pour un même risque α . Elles seront plus faibles pour α , n_1 et n_2 fixés.

Exemple 1 (Un exemple : maintien de l'équilibre selon l'âge).

Est-ce que la capacité à maintenir son équilibre lorsque l'on est concentré est différente selon l'âge² ?

Pour répondre à cette question, $n = 17$ observations ont été recueillies. Des personnes ont été placées sur un plateau mouvant. Elles devaient réagir en appuyant sur un bouton lorsque des signaux arrivaient à intervalles irréguliers. Dans le même temps, elles devaient se maintenir sur le plateau. On a mesuré alors l'amplitude des corrections, d'avant en arrière, effectuées pour rester debout. Les personnes sont subdivisées en 2 groupes : les vieux ($n_1 = 9$) et les jeunes ($n_2 = 8$).

	A	B	C	D	E	F	G	H	I	J	K
1	Numéro	X	groupe		Nombre de X	groupe					
2	1	19	vieux		X	vieux	jeune	F1	F2	Ecart	Abs(Ecart)
3	2	30	vieux		14		2	0.000	0.250	-0.250	0.250
4	3	20	vieux		15		1	0.000	0.375	-0.375	0.375
5	4	19	vieux		17		2	0.000	0.625	-0.625	0.625
6	5	29	vieux		19	2		0.222	0.625	-0.403	0.403
7	6	25	vieux		20	1		0.333	0.625	-0.292	0.292
8	7	21	vieux		21	1	1	0.444	0.750	-0.306	0.306
9	8	24	vieux		22		1	0.444	0.875	-0.431	0.431
10	9	50	vieux		24	1		0.556	0.875	-0.319	0.319
11	10	25	jeune		25	1	1	0.667	1.000	-0.333	0.333
12	11	21	jeune		29	1		0.778	1.000	-0.222	0.222
13	12	17	jeune		30	1		0.889	1.000	-0.111	0.111
14	13	15	jeune		50	1		1.000	1.000	0.000	0.000
15	14	14	jeune		Total						
16	15	14	jeune			9	8			D	0.625
17	16	22	jeune								
18	17	17	jeune								

Fig. 1.5. Feuille de calcul - Test de Kolmogorov-Smirnov

Les calculs ont été réalisés dans le tableur EXCEL (Figure 1.5). Détaillons les étapes :

- Chaque observation est décrite par 2 variables : la variable d'intérêt X et son groupe d'appartenance (colonnes **B** et **C**) ;
- Nous construisons le tableau des effectifs à l'aide de l'outil "Tableaux croisés dynamiques" d'Excel. Nous avons la possibilité de regrouper les données, surtout lorsque les effectifs sont élevés. Ce n'est pas sans dangers néanmoins, des amplitudes d'intervalles mal choisis peuvent masquer les informations pertinentes et fausser les calculs (voir [13], page 145). En règle générale, plus le tableau est détaillé, mieux c'est. En **E**, nous avons les différentes valeurs rencontrées dans le fichier, en **F** (resp. **G**), nous avons les effectifs absolus pour les vieux (resp. jeunes).

² <http://lib.stat.cmu.edu/DASL/Stories/MaintainingBalance.html>

- Nous construisons alors les colonnes des effectifs cumulés, en les ramenant en proportion du total (colonnes **H** et **I**). Nous obtenons ainsi les fonctions de répartition conditionnelles empiriques. Nous noterons au passage que ce dispositif est opérationnel que n_1 soit différent ou égal à n_2 .
- Nous calculons les écarts entre les fonctions de répartition (colonne **J**)
- Et la valeur absolue de l'écart (colonne **K**).
- Nous déduisons alors la statistique du test en récupérant le maximum de cette dernière colonne, soit

$$D = \max(0.250, 0.375, 0.625, \dots, 0.000) = 0.625$$

- La valeur critique du test à 5% est lue dans la table D.1 (page 182), pour $n_1 = 9$ et $n_2 = 8$, nous avons $d_{0.05}(9, 8) = 0.639$
- Conclusion : puisque $D = 0.625 < d_{0.05}(9, 8) = 0.639$, au risque 5%, nos données ne contredisent pas l'hypothèse nulle d'égalité des fonctions de répartition.
- Nous remarquerons néanmoins que nous sommes aux portes de la région critique. Pour un test à 10%, nous rejeterions l'hypothèse nulle.

1.1.2 Approximations pour les grands échantillons

Manipuler les tables statistiques n'est jamais chose aisée. Et elles ne sont pas toujours disponibles. La plupart du temps, les logiciels s'appuient sur des approximations pour calculer les probabilités critiques (les fameuses *p-value*). Elles reposent sur les propriétés asymptotiques des statistiques.

S'agissant du test de Kolmogorov-Smirnov, plusieurs approximations, plus ou moins performantes, sont disponibles. Attention, la convergence est lente, la précision est correcte lorsque n_1 et n_2 prennent des valeurs suffisamment élevées (voir [3], page 110).

Première approximation

Cette approximation est certainement la plus précise. Mais elle est assez complexe à calculer. Nous procédons en deux temps. Nous créons la statistique transformée

$$d = \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}} \times D$$

Pour un test bilatéral, la distribution de loi asymptotique de Kolmogorov-Smirnov (KS) associée s'écrit (voir [3], page 110)

$$P(KS \geq d) = 2 \sum_{j=1}^{+\infty} (-1)^{j+1} \exp(-2j^2 d^2) \quad (1.4)$$

Cette formule nous permet d'obtenir directement la probabilité critique du test bilatéral^{3,4}.

³ Pour un test unilatéral, il suffit de supprimer le facteur 2 devant la somme dans l'équation 1.4

⁴ Une autre expression, moins usitée, peut être utilisée (voir http://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test)

$$P(KS \geq d) = 1 - \frac{\sqrt{2\pi}}{d} \sum_{j=1}^{+\infty} \exp\left(-\frac{(2j-1)^2 \pi^2}{8d^2}\right)$$

Ce type d'expression sont de ceux qui effraient les apprentis statisticiens. La principale idée qu'il faut retenir est que la valeur absolue des items dans la somme tendent très vite vers 0 à mesure que j augmente : concrètement, on peut se contenter de quelques valeurs de j .

Exemple 2 (Application sur l'exemple "Maintien de l'équilibre").

Nous avons obtenu $D = 0.625$. En appliquant la transformation, nous déduisons

$$d = \sqrt{\frac{9 \times 8}{9 + 8}} \times 0.625 = 1.2862$$

	A	B	C
1			
2			
3			
4			
5		D	0.625
6			
7		n1	9
8		n2	8
9			
10		d	1.2862
11			
12			
13			
14		j	Somme
15		1	0.0365592
16		2	-0.0000018
17		3	0.0000000
18			
19		p-value	0.073115

Fig. 1.6. Test de Kolmogorov-Smirnov - 1^{ère} approximation asymptotique

En notant $s_j = (-1)^{j+1} \exp(-2j^2 d^2)$, nous avons :

- $s_1 = (-1)^{1+1} \exp(-2 \times 1^2 \times 1.2862^2) = 0.0365592$;
- $s_2 = -0.0000018$;
- et $s_3 < 0.0000001$

La décroissance est très rapide, on peut s'arrêter dès le troisième terme de la somme pour cet exemple.

Il ne reste plus qu'à calculer la probabilité critique

$$P(KS \geq 1.2862) = 2 \times \sum_{j=1}^3 s_j = 0.073115$$

La p-value est égale à 0.073115. Au risque $\alpha = 5\%$, nous ne pouvons pas rejeter l'hypothèse nulle ; à 10%, oui. Malgré le faible effectif, il n'y a pas d'incohérence avec le test basé sur la loi exacte (section 1). Tous ces calculs sont résumés dans une feuille Excel (Figure 1.6).

Deuxième approximation

Il existe une seconde approximation, équivalente à la précédente si on se limite aux trois premiers termes de la somme. Son intérêt est qu'elle est exprimée différemment, peut être un peu moins complexe (quoique!). Elle est utilisée par plusieurs logiciels statistiques. Nous sommes toujours dans le cadre d'un test bilatéral.

Nous définissons la quantité

$$z = \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}} \times D$$

La probabilité critique p du test est produite en appliquant la règle suivante

$$\begin{aligned} 0 \leq z < 0.27, p &= 1 \\ 0.27 \leq z < 1, p &= 1 - \frac{2.506628}{z}(Q + Q^9 + Q^{25}), \text{ où } Q = \exp(-1.233701 \times z^{-2}) \\ 1 \leq z < 3.1, p &= 2(Q - Q^4 + Q^9 - Q^{16}), \text{ où } Q = \exp(-2 \times z^2) \\ z \geq 3.1, p &= 0 \end{aligned}$$

Exemple 3 (Application sur l'exemple "Maintien de l'équilibre").

A partir toujours de notre exemple, nous calculons

$$z = \sqrt{\frac{9 \times 8}{9 + 8}} \times 0.625 = 1.2862$$

Nous appliquons la formule $p = 2(Q - Q^4 + Q^9 - Q^{16})$, où $Q = \exp(-2 \times z^2)$:

- $Q = \exp(-2 \times 1.2862^2) = 0.036559$;
- Nous pouvons en déduire $Q^4 = 0.00000179$, $Q^9 < 0.00000001$, etc.
- Au final, $p = 2 \times (0.036557) = 0.073115$

Troisième approximation très simplifiée

Il existe une troisième approximation, très simplifiée celle ci. Nous la présentons car de nombreux auteurs y font référence (voir par exemple [13], Table L_{III} , page 352). Elle permet de produire très simplement les seuils critiques associés à différents niveaux de risque α pour les tests bilatéraux (Table 1.1). Il faut que les effectifs n_1 et n_2 soient assez élevés.

Exemple 4 (Application sur l'exemple "Maintien de l'équilibre").

Revenons sur notre exemple de référence. La statistique calculée est égale à $D = 0.625$.

Pour obtenir le seuil critique, nous devons tout d'abord produire la quantité :

α	$d_\alpha(n_1, n_2)$
0.1	$1.22\sqrt{\frac{n_1+n_1}{n_1 \times n_2}}$
0.05	$1.36\sqrt{\frac{n_1+n_1}{n_1 \times n_2}}$
0.025	$1.48\sqrt{\frac{n_1+n_1}{n_1 \times n_2}}$
0.01	$1.63\sqrt{\frac{n_1+n_1}{n_1 \times n_2}}$
0.005	$1.73\sqrt{\frac{n_1+n_1}{n_1 \times n_2}}$
0.001	$1.95\sqrt{\frac{n_1+n_1}{n_1 \times n_2}}$

Tableau 1.1. Tableau des seuils critiques pour le test de Kolmogorov-Smirnov

$$a = \sqrt{\frac{9+8}{9 \times 8}} = 0.4859$$

Au risque $\alpha = 0.05$, le seuil critique approximé sera $d_{0.05}(9, 8) = 1.36 \times 0.4859 = 0.661$. Puisque $D = 0.625 < 0.661$, nos données sont compatibles avec l'hypothèse nulle. Notons la proximité entre le seuil exact (0.639; voir section 1) et le seuil approximé (0.661). La précision n'est pas si mauvaise finalement.

Au risque $\alpha = 0.1$, le seuil à utiliser sera $d_{0.1}(9, 8) = 1.22 \times 0.4859 = 0.593$. Nous rejetons maintenant l'hypothèse d'homogénéité des répartitions.

Approximation pour les tests unilatéraux

Concernant les tests unilatéraux, ils sont également convergents pour les hypothèses alternatives de la forme $H_1 : F_1(X) < F_2(X)$ et $H_1 : F_1(X) > F_2(X)$. La loi de distribution de la statistique de Kolmogorov-Smirnov et la région critique du tests ont définies de la même manière. Par rapport au test bilatéral, seul le seuil critique sera modifié pour le même risque α .

La nouveauté ici est que nous disposons d'une approximation plus facile à manipuler lorsque les effectifs n_1 et n_2 augmentent (voir [13], page 148). La quantité

$$\chi^2 = 4D^2 \frac{n_1 \times n_2}{n_1 + n_2} \quad (1.5)$$

Suit une loi du χ^2 à $ddl = 2$ degrés de liberté. La région critique correspond aux valeurs anormalement élevées de la statistique.

Cette approximation peut être utilisée sur des effectifs réduits. Le test devient conservateur dans ce cas c.-à-d. il a tendance à accepter plus que de raison l'hypothèse nulle. De fait, lorsque nous concluons au rejet de l'hypothèse nulle, nous pouvons être sûr qu'il y a un écart manifestement entre les fonctions de répartition (voir [13], page 150).

Exemple 5 (Comparaison des désaccords).

Cet exemple est intéressant à plus d'un titre : nous mettons en oeuvre un test unilatéral, les effectifs sont suffisamment élevés pour obtenir une approximation satisfaisante, nous travaillons sur les données ordinales.

En effet, le test de Kolmogorov Smirnov est applicable pour les données (au moins) ordinales. Il suffit que l'on puisse classer les observations selon la variable d'intérêt. Dans notre exemple, il s'agit de degré de désaccord des individus par rapport à une proposition qui leur a été faite⁵. Il y a 5 valeurs possibles : totalement d'accord, d'accord, indifférent, désaccord, désaccord fort.

Nous voulons comparer le comportement des hommes ($n_1 = 84$) et des femmes ($n_2 = 91$). Plus précisément, nous souhaitons savoir si les hommes sont plus enclins à être en désaccord par rapport à la proposition qui a été faite. L'hypothèse alternative du test s'écrit donc $H_1 : F_1(X) < F_2(X)$.

Dans la feuille de calcul (Figure 1.7), nous avons les fréquences absolues en colonnes **B** et **C**, avec les effectifs marginaux. Nous en déduisons les fréquences relatives cumulées qui servent à élaborer les fonctions de répartition conditionnelles empiriques (Figure 1.8).

	A	B	C	D	E	F	G	H	I
1		Fréquences absolues		Fréquences cumulées					
2	Modalités	Homme (1)	Femme (2)	Homme (1)	Femme (2)	Ecart (F2-F1)		D-	0.1447
3	Totalement d'accord	10	24	0.119	0.264	0.145		n1	84
4	D'accord	15	15	0.298	0.429	0.131		n2	91
5	Indifférent	19	21	0.524	0.659	0.136			
6	Désaccord	18	17	0.738	0.846	0.108		KHI-2	3.6577
7	Désaccord fort	22	14	1.000	1.000	0.000		ddl	2
8	Total	84	91					p-value	0.1606
9									
10									
11									

Fig. 1.7. Test de Kolmogorov-Smirnov unilatéral sur données ordinales

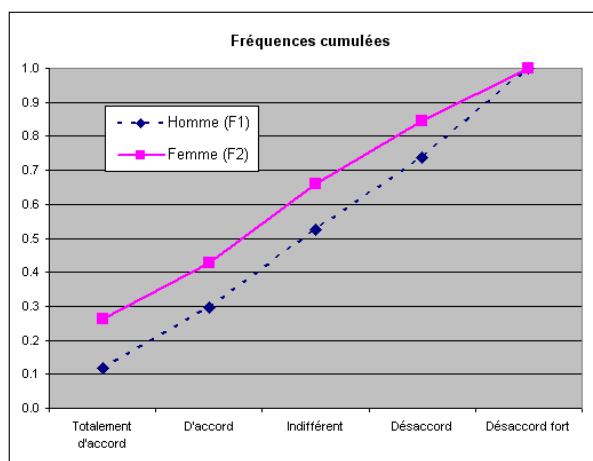


Fig. 1.8. Test de KS unilatéral sur données ordinales - Fonctions de répartition

Dans notre test unilatéral, nous utilisons la statistique

$$D^- = \max_x (F_2(X) - F_1(X))$$

⁵ Voir <http://www.quantitativeskills.com/sisa/statistics/ordhlp.htm#K-S>

Nous obtenons $D^- = 0.1447$. Nous appliquons la formule de transformation

$$\chi^2 = 4 \times 0.1447^2 \times \frac{84 \times 91}{84 + 91} = 3.6577$$

La p-value correspondante pour un $\chi^2(2)$ est $p = 0.1606$. Au risque 5%, nous ne pouvons pas rejeter l'hypothèse nulle d'homogénéité d'opinion des hommes et des femmes face à la proposition qui leur a été faite.

Remarquons que si nous utilisons la version unilatérale de la formule de Kolmogorov-Smirnov (équation 1.4), nous aurions obtenu $p = 0.1599$. L'approximation basée sur la loi du χ^2 est plutôt de bonne facture.

1.1.3 Le cas particulier $n_1 = n_2 = m$

Distribution de la statistique de Kolmogorov-Smirnov

Lorsque les effectifs sont équilibrés, l'expression de la loi de Kolmogorov-Smirnov est simplifiée. Pour le test bilatéral, il s'écrit :

$$P(KS \geq d) = 2 \sum_{j=1}^{[m/k]} (-1)^{j+1} \frac{(m!)^2}{(m-jk)!(m+jk)!} \quad (1.6)$$

pour le test unilatéral

$$P(KS \geq d) = \frac{(m!)^2}{(m-k)!(m+k)!} \quad (1.7)$$

où k est tout entier positif tel que $d = k/m$, $[m/k]$ est la partie entière de la fraction m/k (voir [3], pages 110 et 111 ; voir aussi [1], pages 311 et 312).

Exemple 6 (Comparaison des capacités de mémorisation).

Nous souhaitons comparer les capacités de mémorisation d'élèves en classe de 6-ème et 4-ème. Nous voulons vérifier l'hypothèse : les enfants plus âgés intègrent mieux les informations mémorisées, et font donc moins d'erreur lorsqu'ils doivent les restituer⁶.

On veut vérifier que les enfants du 1^{er} groupe (élèves en quatrième) font moins d'erreur lors de la restitution que ceux du 2nd (élèves en sixième). Il s'agit d'un test unilatéral où l'hypothèse nulle s'écrit $H_1 : F_1(X) > F_2(X)$. Nous allons produire la statistique D^+ .

Les calculs sont résumés dans la feuille EXCEL (figure 1.9) :

⁶ Un explication approfondie est accessible en ligne <http://www.cons-dev.org/elearning/ando/index.html>, il s'agit par ailleurs des données issues de [13], page 146, où on a rajouté une observation supplémentaire de manière à ce que $n_1 = n_2 = m = 10$.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2																
3		X	niveau			Nombre de X	niveau			Fréquence cumulée						
4						X	quatrième	sixième		quatrième	sixième		Ecart		d	0.7
5		39.1	sixième			24.3	1			0.1	0		0.1		m	10
6		41.2	sixième			29.1	1			0.2	0		0.2			
7		45.2	sixième			32.4	1			0.3	0		0.3		k	7
8		46.2	sixième			32.6	1			0.4	0		0.4			
9		48.7	sixième			34.4	1			0.5	0		0.5		A	1.31682E+13
10		48.4	sixième			35.2	1			0.6	0		0.6		B	2.13412E+15
11		40.6	sixième			38.1	1			0.7	0		0.7		p	0.00617
12		52.1	sixième			39.1		1		0.7	0.1		0.6			
13		47.2	sixième			39.2	1			0.8	0.1		0.7			
14		55.0	sixième			40.6		1		0.8	0.2		0.6			
15		35.2	quatrième			40.9	1			0.9	0.2		0.7			
16		39.2	quatrième			41.2		1		0.9	0.3		0.6			
17		40.9	quatrième			41.8	1			1	0.3		0.7			
18		38.1	quatrième			45.2		1		1	0.4		0.6			
19		29.1	quatrième			46.2		1		1	0.5		0.5			
20		34.4	quatrième			47.2		1		1	0.6		0.4			
21		24.3	quatrième			48.4		1		1	0.7		0.3			
22		32.4	quatrième			48.7		1		1	0.8		0.2			
23		32.6	quatrième			52.1		1		1	0.9		0.1			
24		41.8	quatrième			55.0		1		1	1		0			
						Total	10	10					D+			

Fig. 1.9. Test de KS unilatéral sur données équilibrées - Capacité de mémorisation

- La structure de la feuille est classique maintenant pour nous. Nous construisons le tableau croisé dynamique des effectifs (colonnes **G** et **H**) à partir des données (colonnes **B** et **C**). Nous produisons les fréquences relatives cumulées (colonnes **J** et **K**).
- Nous calculons les écarts (colonne **M**), nous recherchons le maximum $d = 0.7$
- $n_1 = n_2 = m = 10$. Puisque $d = 0.7$, un entier k qui convient est $k = 7$, ainsi nous retrouvons l'égalité $d = k/m = 7/10 = 0.7$
- Nous introduisons ces valeurs dans la formule adéquate (équation 1.7). Le numérateur $A = (10!)^2 = 1.31682 \times 10^{13}$, le dénominateur $B = (10 - 7)! \times (10 + 7)! = 2.13412 \times 10^{15}$.
- La probabilité critique du test est

$$p = \frac{A}{B} = \frac{1.31682 \times 10^{13}}{2.13412 \times 10^{15}} = 0.00617$$

- Au risque $\alpha = 5\%$, nous concluons que les élèves plus âgés (qui sont en quatrième) font moins de fautes lors de la restitution des informations mémorisées.

1.1.4 Le test de Kuiper

Le test de Kuiper est une variante du test de Kolmogorov-Smirnov. La statistique du test s'écrit

$$V = D^+ + D^- \quad (1.8)$$

Il est autant sensible aux écarts entre les caractéristiques de tendance centrale (ex. la médiane) qu'aux écarts entre les queues de distribution⁷.

La région critique du test correspond aux grandes valeurs de V . La distribution asymptotique de la statistique est définie par

⁷ http://en.wikipedia.org/wiki/Kuiper's_test

$$P(Kuiper > V) = Q \left(V \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}} \right) \quad (1.9)$$

où

$$Q(\lambda) = 2 \times \sum_{j=1}^{+\infty} (4j^2 \lambda^2 - 1) \times \exp(-2j^2 \lambda^2)$$

En pratique, l'intérêt du test de Kuiper par rapport au test de Kolmogorov-Smirnov reste quand même marginal.

1.2 Le test de Cramer - von Mises

1.2.1 Principe, statistique de test et région critique

Le test de Cramer - von Mises⁸ est une véritable alternative au test de Kolmogorov-Smirnov. Il permet également de tester toute forme de différenciation entre les distributions. Le test d'hypothèses s'écrit

$$H_0 : F_1(X) = F_2(X)$$

$$H_1 : F_1(X) \neq F_2(X)$$

Sa particularité est qu'il exploite différemment les fonctions de répartition empiriques : au lieu de se focaliser sur l'écart maximal, il compile tous les écarts sous la forme d'une somme des carrés des différences. On rapporte généralement qu'il est plus puissant que le test de Kolmogorov-Smirnov, mais la démonstration théorique manque⁹.

La statistique du test s'écrit

$$T = \frac{n_1 \times n_2}{(n_1 + n_2)^2} \sum_{i=1}^{n_1+n_2} [F_1(x_i) - F_2(x_i)]^2 \quad (1.10)$$

Nous pouvons exprimer cette même statistique en passant par les rangs des observations. Soit r_{ik} le rang de l'observation n^oi du sous échantillon Ω_k dans l'ensemble de l'échantillon Ω . La statistique peut être calculée à l'aide de l'expression suivante :

$$T = \frac{U}{n_1 n_2 (n_1 + n_2)} - \frac{4n_1 n_2 - 1}{6(n_1 + n_2)} \quad (1.11)$$

où

⁸ A l'origine, le test permettait de tester l'adéquation d'une fonction de répartition empirique avec une distribution théorique. L'adaptation au test d'homogénéité est due à Anderson (1962).

⁹ Voir http://en.wikipedia.org/wiki/Cramér-von-Mises_criterion

$$U = n_1 \sum_{i=1}^{n_1} (r_{i1} - i)^2 + n_2 \sum_{i=1}^{n_2} (r_{i2} - i)^2 \quad (1.12)$$

La région critique est définie par les valeurs anormalement élevées de T . Les seuils critiques au risque α sont recensées dans des tables statistiques spécifiques¹⁰.

Exemple 7 (Formulation 1 - Activité sportive et indice de masse corporelle).

On souhaite comparer l'indice de masse corporelle (IMC) des lycéens masculins d'une classe de terminale selon le niveau de leur activité sportive : journalier "daily" ($n_1 = 7$ élèves) ou jamais "never" ($n_2 = 7$ élèves). Nous utiliserons la première formulation (équation 1.10).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Cramer - von Mises : Méthode 1												
2														
3														
4		IMC	Sport			Nombre de IMC	Sport			Fréquences cumulées				
5		26.1	DAILY			IMC	DAILY	NEVER		DAILY	NEVER		Ecart	Ecart²
6		23.6	DAILY			22.8	1			0.14285714	0		0.1429	0.0204
7		23.4	DAILY			23		1		0.14285714	0.14285714		0.0000	0.0000
8		30.2	DAILY			23.4	1			0.28571429	0.14285714		0.1429	0.0204
9		22.8	DAILY			23.6	1			0.42857143	0.14285714		0.2857	0.0816
10		24.8	DAILY			23.7	1			0.57142857	0.14285714		0.4286	0.1837
11		23.7	DAILY			24.8	1			0.71428571	0.14285714		0.5714	0.3265
12		26	NEVER			26		1		0.71428571	0.28571429		0.4286	0.1837
13		26.3	NEVER			26.1	1			0.85714286	0.28571429		0.5714	0.3265
14		23	NEVER			26.3		1		0.85714286	0.42857143		0.4286	0.1837
15		33.5	NEVER			27.3		1		0.85714286	0.57142857		0.2857	0.0816
16		28.7	NEVER			28.7		1		0.85714286	0.71428571		0.1429	0.0204
17		27.3	NEVER			30.2	1			1	0.71428571		0.2857	0.0816
18		35.3	NEVER			33.5		1		1	0.85714286		0.1429	0.0204
19						35.3		1		1	1		0.0000	0.0000
20						Total	7	7					Somme	1.5306
21														
22													n1	7
23													n2	7
24														
25													T	0.382653

Fig. 1.10. Test de Cramer - von Mises - Comparaison des indices de masse corporelle (Version 1)

La feuille de calcul est organisée de la manière suivante (Figure 1.10) :

- A partir des données (colonnes **B** et **C**), nous produisons le tableau des fréquences absolues (colonnes **E** à **G**) à l'aide de l'outil "Tableaux croisés dynamiques".
- Nous en déduisons les fréquences relatives cumulées (colonnes **J** et **K**), correspondant aux fonctions de répartition empiriques (Figure 1.11).
- Nous calculons l'écart et le carré de l'écart entre les deux fonctions, pour chaque valeur x_i de X . En cas d'ex-aequo, il faudrait introduire une petite modification dans la formule¹¹.
- Reste à faire la somme $S = 1.5306$, nous obtenons alors la statistique T

$$T = \frac{7 \times 7}{(7 + 7)^2} \times 1.5306 = 0.382653$$

¹⁰ Voir par exemple l'article de T.W. Anderson (1962) , *On the distribution of the two-sample Cramer - von Mises criterion*, Annals of Mathematical Statistics ; accessible en ligne http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177704477, tables 2 à 5

¹¹ Voir la documentation en ligne du logiciel SAS : <http://v8doc.sas.com/sashtml/> ; en particulier la PROC NPAR1WAY dans SAS/STAT

- Pour conclure à la significativité (ou non) de l'écart, nous devons comparer cette valeur avec le seuil critique lue dans la table de Cramer-von Mises (Figure D.2, section D.2). Fait assez inhabituel¹², cette dernière, issue de simulations, ne donne pas les valeurs exactes des seuils pour $\alpha = 10\%$. Nous constatons simplement que notre statistique correspond exactement à la valeur associée à la probabilité critique $\alpha = 0.093240$. C'est la p-value de notre test. De fait, au risque 10%, nous rejetons l'hypothèse nulle d'égalité des fonctions de répartition. Au risque 5%, nous aurions conclu à la compatibilité des données avec H_0 .

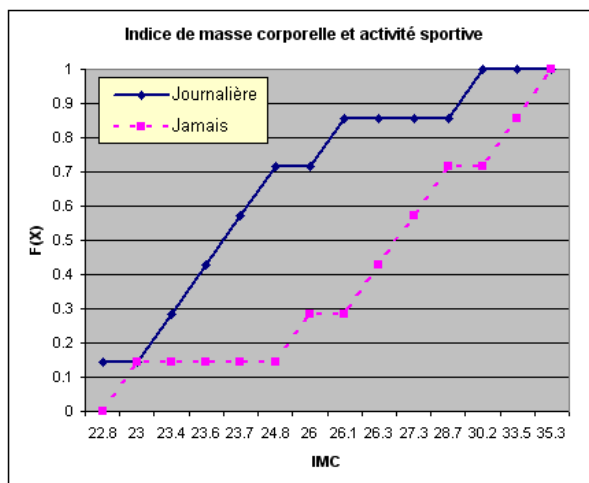


Fig. 1.11. Indice de masse corporelle selon l'activité sportive - Fonctions de répartition

Remarque 3 (Qu'aurait donné le test de Kolmogorov-Smirnov sur le même exemple ?). A titre de curiosité, nous utilisons le test de Kolmogorov-Smirnov sur les mêmes données. A la lumière de notre feuille de calcul (Figure 1.10), le maximum de la valeur absolue de l'écart entre les fonctions de répartitions est égal à $D = 0.5714$. Pour un risque à 10%, nous la confrontons au seuil calculé à l'aide l'approximation asymptotique simplifiée¹³ (table 1.1), soit $d_{0.1}(7, 7) = 1.22\sqrt{\frac{7 \times 7}{7+7}} = 0.6521$. Comme $D = 0.5714 < d_{0.1}(7, 7) = 0.6521$, nous concluons que l'écart entre les fonctions de distribution n'est pas significatif. La conclusion est en contradiction avec celle du test de Cramer - von Mises. Nous dirons aussi, même si c'est un peu subjectif, qu'elle est en contradiction avec *l'impression visuelle*¹⁴ laissée par le graphique des fonctions de répartition conditionnelles (Figure 1.11).

Exemple 8 (Formulation 2 - Activité sportive et indice de masse corporelle).

Nous reprenons le même exemple en utilisation cette fois-ci la seconde formulation (équations 1.11 et 1.12). Cette présentation est très souvent reprise dans les manuels de statistique, mais elle est très rarement illustrée de manière détaillée. C'est ce que nous allons essayer de faire en utilisant toujours un tableur (Figure 1.12) :

¹² On ne comprend dans le sens où l'effectif étant très faible, le nombre de configurations possibles est limité

¹³ Le test exact basé sur la lecture des tables statistiques (ex. [13], table L_{II} , page 350) aboutit à la même conclusion.

¹⁴ En statistique, un graphique bien senti vaut largement autant que des calculs compliqués, ne l'oublions jamais.

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1												
2												
3												
4		IMC	Sport		rang	i		Ecart	Ecart ²		n1	7
5		22.8	DAILY		1	1		0	0		S1	33
6		23.4	DAILY		3	2		1	1			
7		23.6	DAILY		4	3		1	1			
8		23.7	DAILY		5	4		1	1			
9		24.8	DAILY		6	5		1	1			
10		26.1	DAILY		8	6		2	4			
11		30.2	DAILY		12	7		5	25			
12		23	NEVER		2	1		1	1		n2	7
13		26	NEVER		7	2		5	25		S2	232
14		26.3	NEVER		9	3		6	36			
15		27.3	NEVER		10	4		6	36			
16		28.7	NEVER		11	5		6	36		U	1855
17		33.5	NEVER		13	6		7	49			
18		35.3	NEVER		14	7		7	49		T	0.382653

Fig. 1.12. Test de Cramer - von Mises - Comparaison des indices de masse corporelle (Version 2)

- Les données sont organisées en groupes selon la variable SPORT. A l'intérieur de chaque bloc, nous les trions selon la variable d'intérêt IMC (colonnes **R** et **S**). Nous distinguons les observations correspondant à SPORT = DAILY dans la première partie du tableau, SPORT = NEVER dans la seconde.
- Nous créons un second tableau où nous reprenons le rang des observations. Attention, les rangs sont calculés sur la globalité de l'échantillon c.-à-d. la valeur 22.8 a le rang $r_{11} = 1$ dans tout l'échantillon, la seconde valeur 23.4 possède le rang $r_{21} = 3$, etc. Nous faisons de même pour le second sous échantillon c.-à-d. pour la valeur $x_{12} = 23$, nous avons $r_{12} = 2$, pour $x_{22} = 26$, nous obtenons $r_{22} = 7$, etc. Nous plaçons en face des rangs le numéro d'observation dans le sous-échantillon (colonnes **U** et **V**). Tout est prêt maintenant pour calculer la quantité U (équation 1.12).
- Pour la première somme de l'équation 1.12, nous obtenons

$$S_1 = (1 - 1)^2 + (3 - 2)^2 + \dots + (12 - 7)^2 = 33$$

Pour la seconde somme S_2

$$S_2 = (2 - 1)^2 + (7 - 2)^2 + \dots + (14 - 7)^2 = 232$$

Les opérations sont résumées dans les colonnes **X** et **Y** de la feuille de calcul.

- Ainsi, $U = 7 \times 33 + 7 \times 232 = 1855$
- Nous disposons d'un résultat qui permet déjà de porter un jugement sur la significativité de l'écart. En effet, les tables de Cramer-von Mises (Table D.2, section D.2) fournissent quelques probabilités critiques pour les valeurs de la statistique (et les seuils critiques pour un niveau de risque choisi). Nous retrouvons la p-value $p = 0.093240$. Au risque de 10%, nous rejetons l'hypothèse nulle.
- Nous pouvons appliquer la formule (1.11) pour retrouver la valeur de T :

$$T = \frac{1855}{(7 \times 7)(7 + 7)} - \frac{4 \times 7 \times 7 - 1}{6(7 + 7)} = 0.382653$$

Les résultats sont cohérents avec la première approche. C'est toujours rassurant.

1.2.2 Traitement des ex-aequo

Lorsqu'il y a des ex-aequo c.-à-d. 2 individus ou plus prennent la même valeur, la formulation ci-dessus doit être modifiée pour en tenir compte.

Soit G le nombre de valeurs distinctes ($G \leq n$). Pour chaque valeur v_g , nous décomptons le nombre d'observations correspondantes t_g . La statistique de Cramer von Mises s'écrit alors de la manière suivante :

$$T = \frac{n_1 \times n_2}{(n_1 + n_2)^2} \sum_{g=1}^G t_g \times [F_1(v_g) - F_2(v_g)]^2 \quad (1.13)$$

1.3 Lecture des sorties des logiciels

Nous analysons le fichier des demandeurs de crédit (Figure 0.1, 5), largement mis à contribution dans notre support dédié aux tests paramétriques¹⁵. Nous utiliserons dans un premier temps le logiciel libre TANAGRA (<http://eric.univ-lyon2.fr/~ricco/tanagra/>), nous comparerons les résultats avec ceux fournis par la version 9.1 de SAS.

Pour rappel, le fichier de données recense $n = 50$ couples demandeurs de crédit. Plusieurs caractéristiques sont décrites (salaire de l'homme et de la femme, l'acceptation du crédit, le fait d'avoir souscrit à une garantie supplémentaire ou non, le type d'emploi de la personne de référence, etc.).

1.3.1 Comparer les salaires féminins

Dans cette section, nous cherchons à savoir si l'acceptation du crédit est subordonnée au niveau de salaire de la femme. En d'autres termes, on cherche à comparer les fonctions de répartition du salaire selon l'acceptation du crédit (Figure 1.13).

"Acceptation = oui" est la modalité positive. Nous observons sur le graphique : les quantités D^+ et D^- (la hauteur des flèches), les coordonnées en abscisse (valeur du salaire) et en ordonnée (valeur de la fonction de répartition). La statistique de Kolmogorov-Smirnov sera basée sur $D = \max(D^+, D^-)$.

Résultats avec TANAGRA

Le traitement avec TANAGRA (composant **K-S 2-SAMPLE TEST**) nous propose les sorties suivantes¹⁶ (Figure 1.14). Voyons-en le détail :

- TANAGRA indique que la modalité positive (+) est "Acceptation = oui". Il fournit quelques statistiques descriptives concernant la variable d'intérêt "Salaire.Femme", globalement et conditionnellement à "Acceptation". Nous observons notamment que le crédit a été accordé à $n_1 = 34$ couples, $n_2 = 16$ ont vu leur dossier rejeté.

¹⁵ Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf

¹⁶ Pour une description approfondie de la mise en oeuvre de Tanagra sur différents problèmes, voir <http://tutoriels-data-mining.blogspot.com/>

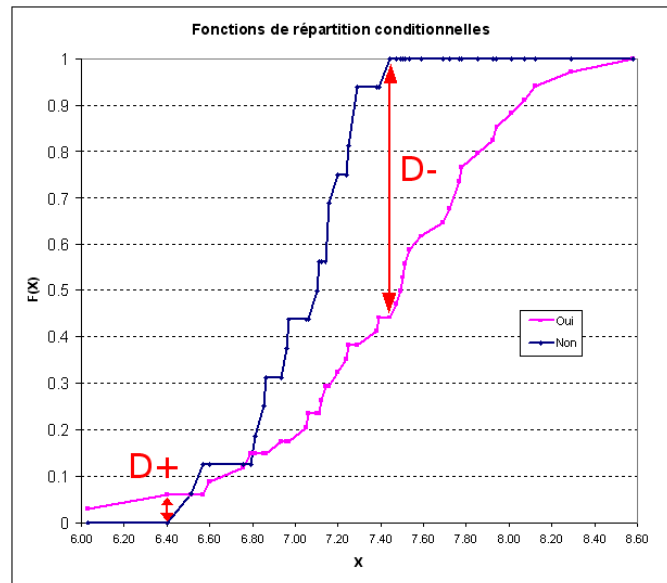


Fig. 1.13. Fonctions de répartition des salaires féminins selon l'acceptation

Results									
Attribute_Y	Attribute_X	Description				Statistical test			
Sal.Femme	Acceptation	Value	Examples	Average	Std-dev	-	Value	Y coord.	EDF coord.
		oui (+)	34	7.4394	0.5483	D+	0.058824	6.4	0.058824
		non (-)	16	7.0331	0.2615	D-	0.558824	7.44	1.000000
		All	50	7.3094	0.5099	Statistics			
							Value	Asymp.Value	p-value
						KS Stat. (D)	0.558824	1.843271	0.002238
						Kuiper's V	0.617647	2.037299	0.007746
						Cramer - von Mises		1.103006	-

Fig. 1.14. Test de Kolmogorov-Smirnov sur le fichier "Crédit" - Sorties de TANAGRA

- La quantité D^+ est défini par l'écart maximal positif entre la fonction de répartition pour "Oui" et pour "Non". Nous constatons que $D^+ = 0.058824$, pour un salaire $X = 6.4$, et lorsque la fonction de répartition prend la valeur $F_1(6.4) = 0.058824$. Ce qui est cohérent avec notre graphique des fonctions de répartition (Figure 1.13).
- De même, pour l'écart positif "non" *versus* "oui", nous obtenons $D^- = 0.558824$, pour $X = 7.44$, avec $F_2(7.44) = 1.0$.
- Seules les quantités D^+ et D^- sont nécessaires par la suite. Les valeurs de X et $F(X)$ sont proposées avant tout pour donner une idée sur le positionnement réciproque des fonctions de répartition.
- La statistique de Kolmogorov-Smirnov est calculée $D = \max(0.058824, 0.558824) = 0.558824$, TANAGRA fournit $d = \sqrt{\frac{34 \times 16}{34 + 16}} \times D = 1.843271$. La probabilité critique du test bilatéral est obtenue à l'aide l'équation¹⁷ 1.4, soit $p = 0.002238$.

¹⁷ Le calcul est limité aux 3 premiers termes de la somme.

- La statistique de Kuiper est également proposée, avec $V = 0.058824 + 0.558824 = 0.617647$. De nouveau, la quantité $\lambda = \sqrt{\frac{34 \times 16}{34 + 16}} \times V = 2.037299$ est formée. La probabilité critique du test est obtenue à l'aide de la formule 1.9, soit $p = 0.007746$
- Enfin, la statistique de Cramer - von Mises est obtenue à l'aide de l'équation 1.10, ou 1.13 s'il y a des ex-aequo, $T = 1.103006$.

Résultats avec SAS

Les mêmes données ont été analysées à l'aide de la procédure **NPARIWAY** du logiciel SAS. Les résultats, bien que présentés sous une forme légèrement différente, sont cohérents avec ceux de TANAGRA, notamment en ce qui concerne les statistiques de test et les probabilités critiques (Figure 1.15, les valeurs à rapprocher avec ceux de TANAGRA sont indiquées à l'aide de petites flèches rouges).

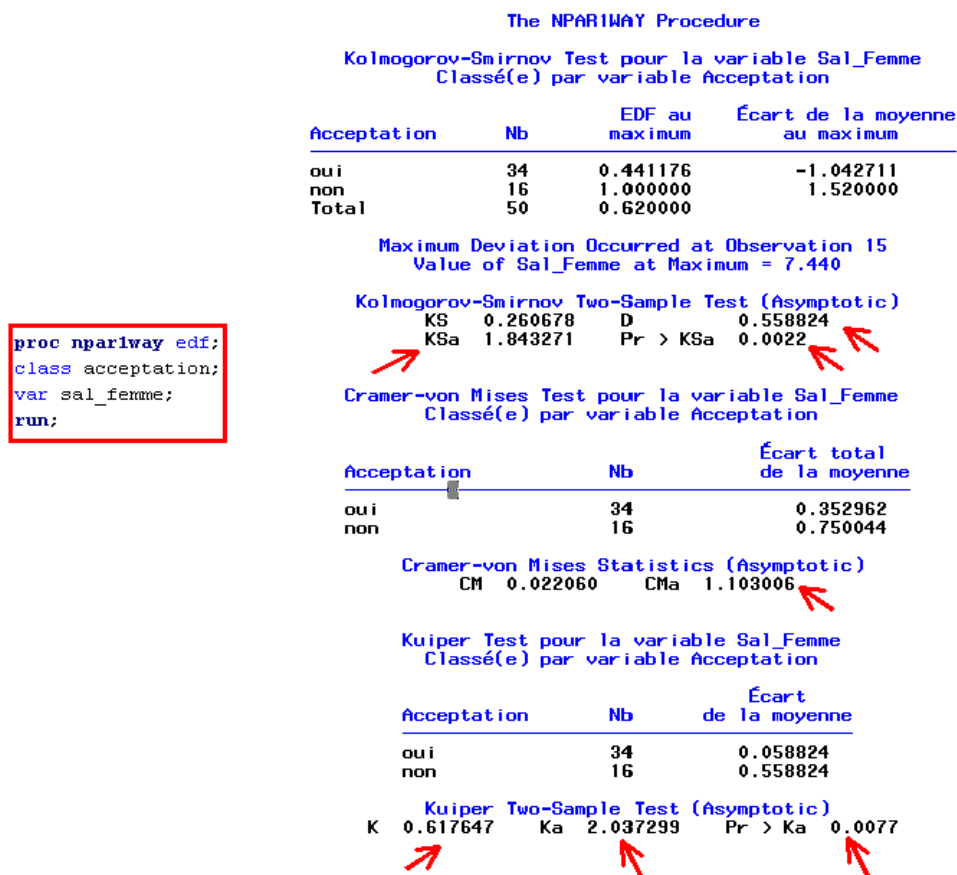
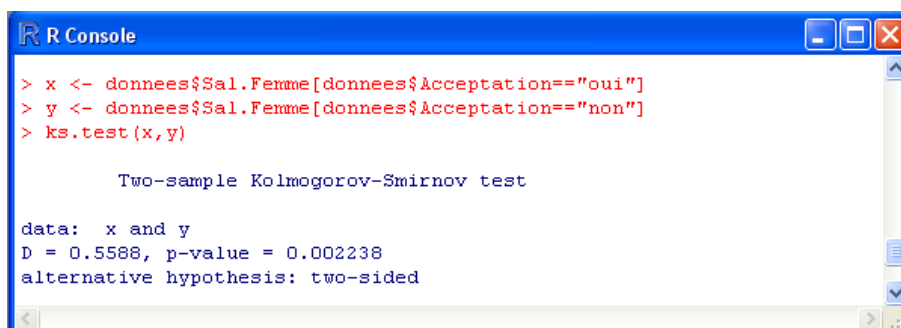


Fig. 1.15. Test de Kolmogorov-Smirnov sur le fichier "Crédit" - Sorties de SAS

Résultats avec R

Avec la commande **ks.test()**, nous pouvons réaliser le test de Kolmogorov-Smirnov avec le logiciel R (un excellent logiciel libre, accessible en ligne - <http://www.r-project.org/>). Fidèle à son habitude, R

propose des sorties très sobres, nous retrouvons la statistique de test $D = 0.5588$ et la probabilité critique associée $p = 0.002238$ (Figure 1.16). Nous observons qu'une petite manipulation préalable des données est nécessaire avant de lancer la procédure, la variable d'intérêt doit être scindée en 2 vecteurs distincts (x, y) .



```

R Console
> x <- donnees$Sal.Femme[donnees$Acceptation=="oui"]
> y <- donnees$Sal.Femme[donnees$Acceptation=="non"]
> ks.test(x, y)

Two-sample Kolmogorov-Smirnov test

data: x and y
D = 0.5588, p-value = 0.002238
alternative hypothesis: two-sided

```

Fig. 1.16. Test de Kolmogorov-Smirnov sur le fichier "Crédit" - Sorties de R

1.3.2 Comparer les salaires masculins

Nous réitérons l'analyse mais avec la variable "Salaire.Homme" cette fois-ci. Nous cherchons à savoir si l'acceptation du crédit est subordonné au salaire de l'homme. Une première approche très simple consiste à élaborer les boîtes à moustaches conditionnelles (Figure 1.17). Il semble qu'il y a une disparité à la fois sur les caractéristiques de tendance centrale (médiane) et les caractéristiques de dispersion (intervalle inter-quartile). Voyons ce que produit le test statistique.

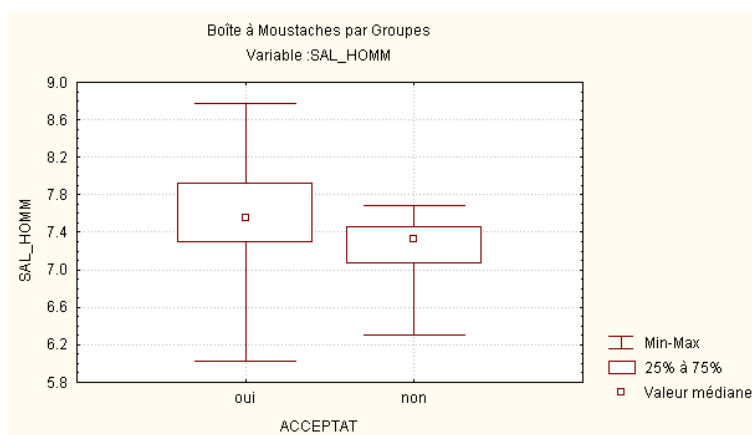


Fig. 1.17. Boîtes à moustaches - Salaire masculin selon l'acceptation - Fichier "Crédit"

Résultats avec TANAGRA

Le test sous Tanagra montre qu'il y a effectivement différenciation des salaires masculins selon l'acceptation (Figure 1.18). La statistique du test est

Results									
Attribute_Y	Attribute_X	Description				Statistical test			
		Value	Examples	Average	Std-dev	-	Value	Y coord.	EDF coord.
Sal.Homme	Acceptation	oui (+)	34	7.5750	0.6155	D+	0.084559	6.83	0.147059
		non (-)	16	7.2281	0.3326	D-	0.466912	7.5	0.937500
		All	50	7.4640	0.5619	Statistics			
							Value	Asymp.Value	p-value
						KS Stat. (D)	0.466912	1.540101	0.017411
						Kuiper's V	0.551471	1.819017	0.032707
						Cramer - von Mises		0.818421	-

Fig. 1.18. TANAGRA - Salaire masculin selon l'acceptation - Fichier "Crédit"

$$D = \max(D^+, D^-) = \max(0.084559, 0.466912) = 0.466912$$

Avec une probabilité critique égale à $p = 0.017411$.

Résultats avec STATISTICA

A l'aide du module "Tests non paramétriques" de STATISTICA (version 5.5). Nous obtenons les résultats sous une forme un peu différente (Figure 1.19). Mais nous retrouvons toutes les informations nécessaires, notamment la statistique de test et la probabilité critique. Pour cette dernière, seul un ordre d'idées est proposé $p < 0.025$.

Test de Kolmogorov-Smirnov (credit approval v4.sta)									
TESTS	par la var. ACCEPTAT								
NON_PARA	Groupe1: 101-non Groupe2: 100-oui								
Variable	Max Nég Différ.	Max Pos Différ.	niveau p	Moyenne non	Moyenne oui	Ec-Type non	Ec-Type oui	N actifs non	N actifs oui
SAL_HOMM	-.466912	.084559	p < .025	7.228126	7.575000	.332550	.615478	16	34

Fig. 1.19. STATISTICA - Salaire masculin selon l'acceptation - Fichier "Crédit"

Remarque 4 (Et par rapport aux tests paramétriques ?). La compatibilité des variables avec la loi normale étant établie par ailleurs (voir [10]), nous avons voulu comparer ces résultats avec le test de comparaison de moyennes d'Aspin-Welch c.-à-d. comparer les moyennes sans faire l'hypothèse d'égalité des variances dans les sous groupes. Nous obtenons (Figure 1.20)

- Pour la comparaison des salaires féminins, : un t de Student de $t = 3.547604$, avec un degré de liberté $ddl = 47.96$ et une probabilité critique $p = 0.00088$
- Pour la comparaison des salaires masculins : un t de Student de $t = 2.581625$, avec un degré de liberté $ddl = 46.92$ et une probabilité critique $p = 0.013012$.

Results						
Attribute_Y	Attribute_X	Description				Statistical test
Sal.Femme	Acceptation	Value	Examples	Average	Std-dev	T
		oui	34	7.4394	0.5483	d.f.
		non	16	7.0331	0.2615	p-value
		All	50	7.3094	0.5099	

Comparaison des salaires féminins

Results						
Attribute_Y	Attribute_X	Description				Statistical test
Sal.Homme	Acceptation	Value	Examples	Average	Std-dev	T
		oui	34	7.5750	0.6155	d.f.
		non	16	7.2281	0.3326	p-value
		All	50	7.4640	0.5619	

Comparaison des salaires masculins

Fig. 1.20. Comparaison des moyennes des salaires l'acceptation - Fichier "Crédit"

Ces résultats sont en concordance avec ceux de Kolmogorov-Smirnov. Nous disposons néanmoins d'une information additionnelle avec ces tests de comparaison de moyennes, il semble que la différenciation repose avant tout sur un écart entre les paramètres de localisation (la moyenne en est une). Nous vérifierons cela dans le chapitre qui vient à l'aide des tests non paramétriques adéquats.

Tests de rang dans un modèle de localisation pour $K = 2$ échantillons

Les tests du chapitre précédent sont destinés à détecter les écarts entre 2 fonctions de répartition sans en préciser la nature. Pour affiner les résultats et comprendre le processus à l'origine de la différenciation entre les sous populations, nous devons avancer d'une étape et nous pencher sur ce mécanisme. Habituellement, deux principales sources de différenciation sont étudiées : les modèles de localisation, ils stipulent que l'écart est attribué au décalage entre les caractéristiques de tendance centrale des distributions que nous confrontons ; les modèles d'échelle, l'écart est due à des dispersions différentes. On ne peut s'empêcher de rapprocher cette classification avec les tests paramétriques usuels, à savoir la comparaison de moyennes et la comparaison des variances lorsque la distribution sous-jacente des données est gaussienne. L'avantage des tests non paramétriques est que cette contrainte de normalité est levée, le champ d'application des procédures statistiques est plus large. **Nous supposons néanmoins que la distribution sous-jacente des données est continue.**

Dans ce chapitre, nous étudierons le modèle de localisation. Nous comparons des caractéristiques de tendance centrale, pas nécessairement la moyenne, surtout pas d'ailleurs. On entend généralement par paramètre de localisation un estimateur robuste de tendance centrale. "Robuste" s'entend par "tolérance aux valeurs extrêmes (atypiques)". La médiane en fait partie, nous pouvons l'utiliser explicitement pour différencier les distributions (section 2.5). Mais il en existe d'autres (voir [3], pages 53 à 80).

Formellement, le test d'hypothèses sera ré-écrit de la manière suivante (voir [1], page 316) :

$$\begin{aligned}H_0 : F_1(X) &= F_2(X + \theta), \quad \theta = 0 \\H_1 : F_1(X) &= F_2(X + \theta), \quad \theta \neq 0\end{aligned}$$

θ est le **paramètre de translation**. Il traduit le décalage entre les fonctions de répartition. Bien entendu, il est possible de spécifier des tests unilatéraux. Pour traduire l'idée " X prend stochastiquement des valeurs plus faibles (resp. plus élevées) dans le premier échantillon, nous précisons $H_1 : \theta < 0$ (resp. $H_1 : \theta > 0$). Dans la figure 2.1, nous observons les fonctions de répartition à comparer sur des données simulées, le décalage est matérialisé par le paramètre θ . Attention, de la manière avec laquelle nous avons ré-écrit le test d'hypothèses, l'écart est censé être constant tout au long des fonctions de répartition. Nous pouvons faire de l'inférence sur θ , l'estimer à partir des données et calculer les intervalles de confiance (voir [3], pages 158 à 169).

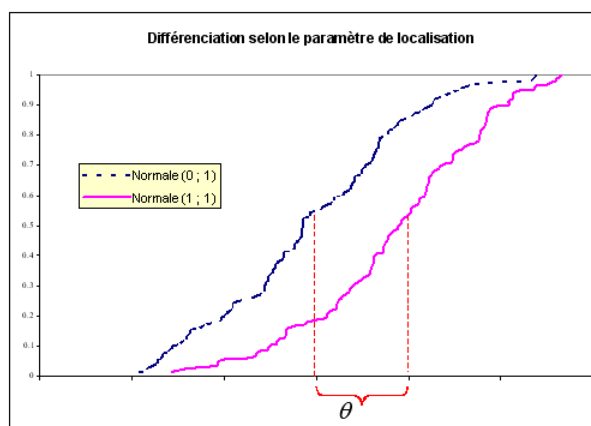


Fig. 2.1. Paramètre de translation - Décalage entre 2 fonctions de répartition

Notons qu'il existe d'autres manières d'appréhender le décalage systématique entre les fonctions de répartition : en termes de comparaison directe des paramètres de localisation, "la médiane dans la première sous population est-elle supérieure à celle de la seconde?" ; en termes de probabilistes, "la probabilité que la valeur mesurée dans la première sous population soit plus grande que celle mesurée dans la seconde est-elle supérieure à 0.5?" (voir [13], page 129).

2.1 Test de Wilcoxon - Mann - Whitney

Le test de Wilcoxon - Mann - Whitney est certainement le plus populaire des tests non paramétriques¹. Il recouvre en réalité 2 formulations, totalement équivalentes (ils peuvent se déduire l'un de l'autre), le test de Wilcoxon d'une part, le test de Mann-Whitney d'autre part. Nous nous concentrerons sur ce dernier dans un premier temps.

2.1.1 Rangs, somme des rangs et traitement des ex-aequos

La très grande majorité des tests non paramétriques reposent sur la notion de rangs². L'idée est de substituer aux valeurs leur numéro d'ordre *dans l'ensemble des données*.

Rangs et somme des rangs

A la valeur x_i de l'individu $n^o i$ correspond maintenant son rang r_i . Si nous souhaitons spécifier l'appartenance au groupe, à la valeur x_{ik} correspond le rang r_{ik} . Mais attention, le rang est toujours

¹ Pour une représentation détaillée et récente de cette technique, voir http://www.tqmp.org/doc/vol4-1/p13-20_Nachar.pdf. Au passage, signalons que les articles de la revue *Tutorials in quantitative methods for Psychology* sont accessibles en ligne - <http://www.tqmp.org/content.php>

² Nous étudierons plus loin les caractéristiques génériques des statistiques basées sur les rangs (voir chapitre 5).

calculé en référence à l'ensemble de l'échantillon Ω . Nous pouvons dès lors former la somme des rangs pour chaque sous échantillon :

$$S_k = \sum_{i=1}^{n_k} r_{ik} \quad (2.1)$$

De manière pragmatique, cette transformation introduit 2 conséquences importantes : (a) la distribution des données (les rangs) devient nécessairement symétrique, quelle qu'ait été la distribution initiale des valeurs ; (b) le rôle des points atypiques est considérablement amoindri.

Numéro global	Numéro dans le groupe	IMC	Sport	Rang
1	1	22.8	DAILY	1
2	2	23.4	DAILY	3
3	3	23.6	DAILY	4
4	4	23.7	DAILY	5
5	5	24.8	DAILY	6
6	6	26.1	DAILY	8
7	7	30.2	DAILY	12
8	1	23	NEVER	2
9	2	26	NEVER	7
10	3	26.3	NEVER	9
11	4	27.3	NEVER	10
12	5	28.7	NEVER	11
13	6	33.5	NEVER	13
14	7	35.3	NEVER	14

n1	7
S1	39
r_barre_1	5.571

n2	7
S2	66
r_barre_2	9.429

Fig. 2.2. Calcul des rangs - IMC selon l'activité sportive

Prenons un petit exemple pour détailler le processus. Dans le tableau des données des comparaisons des indices de masses corporelles selon l'activité sportive (Figure 2.2). Nous avons attribué un numéro global à chaque observation pour plus de clarté. À partir des valeurs de la variable d'intérêt (IMC), nous attribuons le rang. Ainsi, à la valeur $x_1 = 22.8$, qui est la plus petite, est associée le rang $r_1 = 1$; à la valeur $x_2 = 23.4$ est associée le rang $r_2 = 3$, c'est la troisième plus petite valeur de l'échantillon ; etc.

Pour calculer la somme des rangs dans le groupe $n^o 1$ c.-à-d. "Sport = Daily", nous formons :

$$S_1 = \sum_{i=1}^{n_1=7} r_{i1} = 1 + 3 + 4 + 5 + 6 + 8 + 12 = 39$$

Nous pouvons calculer le rang moyen, il indique la localisation du groupe

$$\bar{r}_1 = \frac{S_1}{n_1} = \frac{39}{7} = 5.571$$

De la même manière, nous aurons $n_2 = 7$, $S_2 = 66$ et $\bar{r}_2 = 9.429$.

En moyenne, les observations associées au premier groupe sont plutôt situées à gauche de celles du second groupe. Nous pourrions également raisonner en termes de "tendance centrale" ou de "localisation" des distributions.

Traitement des ex-aequos - Le principe des rangs moyens

Lorsqu'il y a des ex-aequos dans les valeurs, deux approches sont possibles. La méthode des rangs aléatoires attribue aléatoirement les rangs aux observations confondues. Dans ce cas, aucune modification des tables et lois asymptotiques existantes n'est nécessaire. Cependant, d'une part, la puissance du test est plus faible que celle de la méthode que nous présentons plus bas ; d'autre part, la possibilité que la conclusion du test puisse être différente d'un coup sur l'autre³ (selon l'attribution des rangs) n'est pas très défendable, cela arrive lorsque la statistique observée est située à la lisière de la région critique.

La méthodes des rangs moyens procède de la manière suivante : les observations possédant des valeurs identiques se voient attribuer la moyenne de leurs rangs. Cette approche est plus puissante. Les statistiques de test (plus précisément leur variance) sont cependant modifiées. Nous préciserons la nature de la correction à introduire lors du calcul des lois et des statistiques lorsque nous présenterons les différents tests qui composent ce chapitre.

Explication orale des maladies	Niveau anxiété	Rang brut	Rang moyen
absent	6	1	1.5
present	6	2	1.5
absent	7	3	5
absent	7	4	5
absent	7	5	5
absent	7	6	5
absent	7	7	5
absent	8	8	9.5
absent	8	9	9.5
present	8	10	9.5
present	8	11	9.5
absent	9	12	12
absent	10	13	16
absent	10	14	16
absent	10	15	16
absent	10	16	16
present	10	17	16
present	10	18	16
present	10	19	16
present	11	20	20.5
present	11	21	20.5
absent	12	22	24.5
absent	12	23	24.5
present	12	24	24.5
present	12	25	24.5
present	12	26	24.5
present	12	27	24.5
absent	13	28	29.5
present	13	29	29.5
present	13	30	29.5
present	13	31	29.5
present	14	32	33
present	14	33	33
present	14	34	33
present	15	35	36
present	15	36	36
present	15	37	36
present	16	38	38
present	17	39	39

absent	
n1	16
S1	200
r_barre_1	12.50

present	
n2	23
S2	580
r_barre_2	25.22

Fig. 2.3. Calcul des rangs moyens en cas d'ex-aequo

Pour l'heure, illustrons la méthodes des rangs moyens sur un petit exemple (Figure 2.3). La variable d'intérêt X correspond au niveau d'anxiété d'enfants face à la socialisation orale dans des sociétés primitives. On oppose le groupe de $n_1 = 16$ enfants issus d'une société où la tradition orale expliquant les

³ L'utilisateur aura vite fait de remettre en cause l'intégrité du logiciel.

effets des maladies est "absente", avec celui où elle est "présente" $n_2 = 23$. Comme il s'agit d'une échelle (niveau d'anxiété), il y a bien entendu de nombreux ex-aequo (voir [13], pages 125 à 128 et 132 à 136).

Les observations ont été triées selon les valeurs de X croissantes. Un numéro global sert à repérer les individus. Il correspond aux rangs bruts. Il ne tient pas compte des ex-aequo. Puis, dans un deuxième temps, pour des observations ayant des valeurs identiques, nous attribuons la moyenne des rangs associés. Par exemple, les deux premières plus petites observations présentent la même valeur $x_1 = x_2 = 6$, nous leur attribuons le rang moyen $r'_1 = r'_2 = \frac{1+2}{2} = 1.5$; pour les observations $x_3 = x_4 = x_5 = x_6 = x_7 = 7$, nous produisons le rang $r'_3 = r'_4 = r'_5 = r'_6 = r'_7 = \frac{3+4+5+6+7}{5} = 5$; etc.

La somme et la moyenne des rangs conditionnellement aux groupes sont calculées sur les rangs moyens, soit :

- le groupe n^o1 (tradition absente) : $n_1 = 16$, $S_1 = 200$ et $\bar{r}_1 = 12.5$
- le groupe n^o2 (tradition présente) : $n_2 = 23$, $S_2 = 580$ et $\bar{r}_2 = 25.22$

Si l'on se réfère aux rangs moyens. Le niveau d'anxiété semble plus faible dans le 1^{er} groupe, celui des enfants non informés des problèmes liés à la maladie (tradition orale = absente). Il reste à confirmer cela statistiquement.

2.1.2 Statistiques de rang linéaires

Le test que Wilcoxon - Mann - Whitney, comme la grande majorité des tests basés sur les rangs de ce chapitre, s'appuie sur une statistique de rang linéaire. Admettons, sans restreindre la portée du discours, que le 1^{er} groupe serve de groupe de référence, la statistique de rang linéaire s'écrit (voir [1], page 316; [3], page 106) :

$$T = \sum_{i=1}^{n_1} f(r_{i1}) \quad (2.2)$$

où $s_i = f(r_i)$ s'appelle **code** ou **score**, et $f(\cdot)$ est une **fonction score**.

On retrouve souvent une autre écriture équivalente

$$T = \sum_{i=1}^n c_i \times f(r_i) \quad (2.3)$$

où c_i est une fonction indicatrice qui vaut 1 lorsque l'individu n^oi appartient à Ω_1 , 0 sinon.

La statistique T possède une propriété très importante, elle converge très rapidement vers la loi normale lorsque la taille des sous échantillons augmentent. Nous mettrons largement à profit cette propriété.

Le choix de la fonctions score $f(\cdot)$ dépend des informations que l'on souhaite mettre en avant. Il peut également reposer sur les caractéristiques des données sous-jacentes aux rangs (la variable initiale X). De fait, si le choix est judicieux, les performances sont grandement améliorées, s'approchant même de celles des statistiques paramétriques équivalentes. Par exemple, avec les *codes normaux*, nous détaillerons cela plus loin, le test associé est aussi puissant que le test de Student de comparaison de moyennes, même lorsque la distribution des données est gaussienne.

Dans cette section, s'agissant du test de Wilcoxon - Mann - Whitney, la fonction $f(\cdot)$ correspond à la fonction identité, les codes correspondent directement aux rangs bruts.

2.1.3 Statistique de test et région critique

Test bilatéral

La statistique de Mann et Whitney utilise la somme des rangs. Nous retrouvons bien l'idée de décalage entre les distributions basé sur leurs localisations respectives. Pour le test bilatéral

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

Nous calculons les quantités

$$U_1 = S_1 - \frac{n_1(n_1 + 1)}{2} \quad (2.4)$$

$$U_2 = S_2 - \frac{n_2(n_2 + 1)}{2} \quad (2.5)$$

Par convention, la statistique de Mann Whitney correspond à la plus petite quantité, soit

$$U = \min(U_1, U_2) \quad (2.6)$$

Lorsque l'hypothèse nulle est vraie, l'espérance et la variance de U s'écrivent :

$$E(U) = \frac{1}{2}n_1n_2 \quad (2.7)$$

$$V(U) = \frac{1}{12}(n_1 + n_2 + 1)n_1n_2 \quad (2.8)$$

La région critique du test correspond aux valeurs exagérément élevées ou exagérément faibles de U par rapport à son espérance.

Exemple 9 (Comparaison des IMC selon l'activité sportive).

Reprenons notre exemple des lycéens plus ou moins sportifs (figure 2.2). Nous complétons les calculs avec les quantités U_1 , U_2 et U :

$$- U_1 = 39 - \frac{7(7+1)}{2} = 11$$

$$- U_2 = 66 - \frac{7(7+1)}{2} = 38$$

$$- U = \min(U_1, U_2) = \min(11, 38) = 11$$

- Nous devrions comparer cette valeur observée avec les valeurs critiques lues dans la table si nous nous référons au schéma classique. Mais cette dernière est organisée un peu différente dans ce support (section D.3), elle nous fournit la probabilité $P(MW \leq U)$. Nous pouvons donc obtenir directement la probabilité critique du test en calculant $p = 2 \times P(MW \leq U)$. Dans notre cas, la table nous indique $P(MW \leq 11) = 0.049$, nous pouvons en déduire $p = 2 \times 0.049 = 0.098$. Au seuil 5%, nous ne pouvons pas rejeter l'hypothèse d'égalité de l'amplitude des réactions des individus selon leur groupe d'appartenance (jeunes ou personnes âgées).

- Au seuil 10%, nous rejeterions l'hypothèse nulle.

- Nous remarquerons que les résultats sont cohérents avec ceux proposés par le test de Kolmogorov-Smirnov sur les mêmes données.

2.1.4 Loi asymptotique - L'approximation normale

Lorsque les échantillons atteignent une taille suffisamment élevés ($n_1 > 8$ et $n_2 > 8$ selon [1], page 317; [11], page 343)⁴. La distribution de la statistique U converge vers la loi normale de moyenne $E(U)$ et de variance $V(U)$.

La statistique centrée et réduite

Pour un test bilatéral, nous pouvons donc définir la statistique centrée réduite

$$Z = \frac{U - \frac{1}{2}n_1n_2}{\sqrt{\frac{1}{12}(n_1 + n_2 + 1)n_1n_2}} \quad (2.9)$$

La région critique du test au niveau de signification α

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

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

Exemple 10 (Comparaison des IMC selon l'activité sportive).

Reprenons notre exemple ci-dessus (figure 2.2). La démarche est un peu sujette à caution ici, les effectifs sont faibles, à la lisière des valeurs suffisantes pour une bonne approximation. Nous la mettons en oeuvre néanmoins, ne serait-ce que pour pouvoir comparer les résultats avec ceux de la statistique exacte. Nous complétons la feuille de calcul en formant la statistique centrée et réduite Z (Figure 2.4) :

- Nous calculons $E(U) = \frac{1}{2} \times 7 \times 7 = 24.5$
- Puis $V(U) = \frac{1}{12} \times (7 + 7 + 1) \times 7 \times 7 = 61.2500$
- Nous formons Z avec

$$Z = \frac{11 - 24.5}{\sqrt{61.2500}} = -1.7250$$

- Au risque $\alpha = 5\%$, nous constatons que $|Z| = 1.7250 < 1.96 = u_{0.975}$, $u_{0.975}$ est le seuil critique, quantile d'ordre 0.975 de la loi normale centrée réduite. Les données sont compatibles avec l'hypothèse nulle d'égalité des IMC dans les groupes.
- Nous pouvons également calculer directement la probabilité critique, elle est égale à $p = 0.0845$. La valeur est plus ou moins éloignée de la probabilité critique *exacte* calculée à partir des tables de Mann et Whitney (qui était égale à 0.098, rappelons-le). Il reste toutefois que les conclusions sont identiques : acceptation de H_0 à 5%, rejet à 10%.

⁴ $n_1 > 10$ et $n_2 > 10$, ou $n_1 > 12$ et $n_2 = 3$ ou 4 selon [13], page 132; $n_1 + n_2 > 20$ avec $n_1 > 3$ et $n_2 > 3$ selon [2], page 19; etc. Bref, de l'ordre d'un dizaine d'observations dans les groupes.

Numéro global	Numéro dans le groupe	IMC	Sport	Rang
1	1	22.8	DAILY	1
2	2	23.4	DAILY	3
3	3	23.6	DAILY	4
4	4	23.7	DAILY	5
5	5	24.8	DAILY	6
6	6	26.1	DAILY	8
7	7	30.2	DAILY	12
8	1	23	NEVER	2
9	2	26	NEVER	7
10	3	26.3	NEVER	9
11	4	27.3	NEVER	10
12	5	28.7	NEVER	11
13	6	33.5	NEVER	13
14	7	35.3	NEVER	14

n1	7
S1	39
r_barre_1	5.571

n2	7
S2	66
r_barre_2	9.429

U1	11
U2	38

U	11
---	----

E(U)	24.5000
V(U)	61.2500

Z	1.7250
---	--------

u_0.975	1.9600
---------	--------

p	0.0845
---	--------

Avec correction de continuité	
Z	1.6611
p	0.0967

Fig. 2.4. Mann et Whitney - Comparaison des IMC selon l'activité sportive

La correction de continuité

Lorsque les effectifs sont de taille modérée (c'est le cas pour notre exemple), nous pouvons améliorer l'approximation normale en introduisant la correction de continuité. Notre statistique de test s'écrit pour un test bilatéral :

$$|Z| = \frac{|U - E(U)| - 0.5}{\sqrt{V(U)}} \quad (2.10)$$

La règle de décision n'est pas modifiée.

Reprenons notre exemple (Figure 2.4). Nous calculons

$$|Z| = \frac{|11 - 24.5| - 0.5}{\sqrt{61.2500}} = 1.6611$$

Avec une probabilité critique $p = 0.0967$. Manifestement, nous nous rapprochons de la *bonne* valeur (0.098), l'approximation est plus précise.

Lorsque les effectifs sont élevés, plusieurs dizaines d'observations, la correction est négligeable. Notons que les logiciels diffèrent sur la manière d'introduire la correction dans les calculs. Certains l'introduisent d'office, d'autres non, d'autres encore donnent à l'utilisateur la possibilité de choisir. Il nous faut en tous les cas vérifier cette option lors de la lecture des résultats.

Le cas des tests unilatéraux

Les tests unilatéraux ($\theta < 0$ ou $\theta > 0$) ne posent pas de difficultés particulières, il faut simplement introduire correctement la correction de continuité. La statistique générique s'écrit (voir [13], page 132 ; la formule correspond à la statistique de Wilcoxon, mais le principe est le même) :

$$Z = \frac{U - E(U) \pm 0.5}{\sqrt{V(U)}} \quad (2.11)$$

Pour un test unilatéral à gauche, nous rajouterons la valeur 0.5 ; à droite, nous la retrancherons.

La région critique pour un test unilatéral à gauche (resp. à droite) au risque α devient *R.C.* : $Z \leq u_\alpha$ (resp. *R.C.* : $Z \geq u_{1-\alpha}$).

Exemple 11 (Comparaison des IMC selon l'activité sportive).

Un médecin du sport vient nous dire qu'a priori, l'activité sportive engendre des lycéens plus minces. A son sens, on devrait donc d'emblée introduire l'hypothèse alternative $H_1 : \theta < 0$ c.-à-d. IMC est plus faible dans le groupe n°1, celui des personnes ayant une activité sportive journalière, nous avons un test unilatéral à gauche.

Nous disposons déjà de toutes les informations pour former le test dans la feuille de Excel précédente (Figure 2.4) :

- Nous formons

$$Z = \frac{(11 - 24.5) + 0.5}{\sqrt{61.2500}} = -1.6611$$

- La région critique pour le test unilatéral devient *R.C.* : $Z \leq u_\alpha$. Pour un test à 5%, nous avons $u_{0.05} = -1.6449$.
- Au risque 5%, puisque $Z = -1.6611 < -1.6449 = u_{0.05}$, nous concluons au rejet de l'hypothèse nulle, les sportifs ("Sport = Daily") sont généralement moins gros avec un indice de masse corporelle plus faible.

Pour rappel, avec $U = 11$ pour un test unilatéral, la table de Mann et Whitney (section D.3) nous fournit la probabilité critique $p = 0.049$. Les conclusions du test exact et du test basé sur l'approximation normale (avec correction de continuité) sont cohérentes. Au risque 5%, nous sommes à la lisière de la région critique.

2.1.5 Correction pour les ex-aequo

Les tests non paramétriques ne reposent pas sur des hypothèses concernant la distribution sous-jacente des données. En revanche, elle suppose la continuité de la fonction de répartition. Or, en pratique, les données sont mesurées avec une certaine précision. Personne n'ira mesurer l'âge en secondes, et il y a forcément des ex-aequo si on la mesure en années (dès que l'on est obligé d'arrondir les valeurs). De même, lorsque les variables sont ordinales par nature (ex. des notes, des degrés de préférence), les ex-aequo sont inhérents au type des données (voir [2], page 16 et 17).

Lorsque les ex-aequo sont associés à des individus du même groupe, la statistique du test n'est pas modifiée. En revanche, lorsque des individus de groupes différents présentent la même valeur et se voient donc attribués des rangs (moyens) identiques, la statistique du test est modifiée par rapport à la méthode des rangs aléatoires. Néanmoins la différence est négligeable.

En revanche, la variance de la statistique doit être corrigée. Ainsi, lorsque nous voulons utiliser l'approximation normale pour définir la région critique du test, nous devons utiliser la formule suivante (voir [13], page 134, pour la variance de la statistique de Wilcoxon, c'est la même pour la statistique de Mann-Whitney ; [1], page 325 ; [3], pages 132 et 133) :

$$\tilde{V}(U) = V(U) \times \left(1 - \frac{\sum_{g=1}^G t_g(t_g^2 - 1)}{n^3 - n} \right) \quad (2.12)$$

Où $n = n_1 + n_2$, l'effectif total, G est le nombre de valeurs distinctes dans l'échantillon Ω , t_g est le nombre d'observation associée à la valeur $n^o g$.

Étonnamment, dans la grande majorité des références traitant des tests non paramétriques (mis à part les ouvrages cités ci-dessus), cette correction est peu mise en avant. Pourtant elle peut être assez importante, surtout lorsque le nombre de valeurs différentes est faible, induisant un grand nombre d'ex-aequo. C'est immanquablement le cas lorsque les variables sont ordinales. Les principaux logiciels de statistiques utilisent systématiquement la correction pour ex-aequo lors du calcul de la statistique centrée réduite pour l'approximation normale.

Exemple 12 (Tester le degré d'anxiété).

Reprenons notre exemple de degré d'anxiété des enfants (Figure 2.3). Nous complétons la feuille de calcul pour déjà construire la statistique du test bilatéral. Nous y insérons les deux estimations de la variance, sans et avec la correction pour ex-aequo⁵.

Voici dans un premier temps la description du calcul de U , et $|Z|$ sans la correction pour ex aequo (Figure 2.5) :

- A partir de la somme des rangs, nous calculons $U_1 = 200 - \frac{16(16+1)}{2} = 64$, $U_2 = 580 - \frac{23(23+1)}{2} = 304$.
Nous en déduisons $U = \min(64, 304) = 64$
- L'espérance $E(U) = \frac{16 \times 23}{2} = 184$.
- La variance $V(U) = \frac{(16+23+1) \times 16 \times 23}{12} = 1226.6667$
- La statistique $|Z|$ pour le test bilatéral s'obtient par le rapport (sans la correction de continuité)

$$|Z| = \frac{|64 - 184|}{\sqrt{1226.6667}} = 3.4262$$

- La p-value du test est $p = 0.00061$. Au risque 5%, nous concluons à une différence significative entre les niveaux d'anxiété des enfants.

⁵ Parce que les effectifs sont *assez* élevés, et surtout pour ne pas interférer avec le sujet de cette section, nous n'introduirons pas la correction de continuité dans notre exemple.

Tradition orale	Niveau_Anxiete	Rang brut	Rang moyen
absent	6	1	1.5
present	6	2	1.5
absent	7	3	5
absent	7	4	5
absent	7	5	5
absent	7	6	5
absent	7	7	5
absent	8	8	9.5
absent	8	9	9.5
present	8	10	9.5
present	8	11	9.5
absent	9	12	12
absent	10	13	16
absent	10	14	16
absent	10	15	16
absent	10	16	16
present	10	17	16
present	10	18	16
present	10	19	16
present	11	20	20.5
present	11	21	20.5
absent	12	22	24.5
absent	12	23	24.5
present	12	24	24.5
present	12	25	24.5
present	12	26	24.5
present	12	27	24.5
absent	13	28	29.5
present	13	29	29.5
present	13	30	29.5
present	13	31	29.5
present	14	32	33
present	14	33	33
present	14	34	33
present	15	35	36
present	15	36	36
present	15	37	36
present	16	38	38
present	17	39	39

absent
n1 16
S1 200
r_barre_1 12.50

present
n2 23
S2 580
r_barre_2 25.22

U1 64
U2 304
U 64
E(U) 184

Sans tenir compte des ex-aequo
V(U) 1226.6667
Z 3.4262
p-value 0.00061

En tenant compte des ex-aequo
n 39
Facteur correction 0.9857
V~(U) 1209.1606
Z 3.4510
p-value 0.00056

Tableau des valeurs uniques				
g	Valeur	Rang associé	t_g	tg(tg^2-1)
1	6	1.5	2	6
2	7	5	5	120
3	8	9.5	4	60
4	9	12	1	0
5	10	16	7	336
6	11	20.5	2	6
7	12	24.5	6	210
8	13	29.5	4	60
9	14	33	3	24
10	15	36	3	24
11	16	38	1	0
12	17	39	1	0
Somme				846

Fig. 2.5. Mann et Whitney - Comparaison de l'anxiété - Corrections pour ex-aequo

Pour introduire la correction pour les ex-aequo, nous devons calculer le nombre de valeurs uniques différentes G et les effectifs correspondants (t_g). C'est le rôle du tableau "Tableau des valeurs uniques". Les valeurs que l'on retrouve dans le fichier sont $v_1 = 6, v_2 = 7, v_3 = 8, \dots, v_{12} = 17$, soit $G = 12$. Nous comptabilisons les effectifs associés $t_1 = 2, t_2 = 5, \dots, t_{12} = 1$. Nous pouvons dès lors former la somme

$$\sum_{g=1}^{G=12} t_g(t_g^2 - 1) = 6 + 120 + 60 + 0 + 336 + \dots + 0 = 846$$

Sachant que $n = 16 + 23 = 39$, la nouvelle variance, en introduisant le facteur de correction tenant compte des ex-aequo, devient

$$\begin{aligned} \tilde{V}(U) &= V(U) \times \left(1 - \frac{\sum_{g=1}^G t_g(t_g^2 - 1)}{n^3 - n} \right) \\ &= 1226.6667 \times \left(1 - \frac{846}{39^3 - 39} \right) \\ &= 1209.1606 \end{aligned}$$

Dès lors $|Z| = \frac{|64-184|}{\sqrt{1209.1606}} = 3.4510$ et la probabilité critique du test est $p = 0.00056$.

La correction est assez faible. Ce qui est sûr en tous les cas, c'est qu'introduire la correction pour valeurs ex-aequo ne peut que réduire la variance estimée. La statistique centrée et réduite est à l'inverse augmentée. Les résultats seront toujours un peu plus significatifs. De fait, ne pas tenir compte des ex-aequo correspond à un comportement conservateur, on favorise l'acceptation de l'hypothèse nulle (voir [13], page 136).

2.1.6 La variante de Wilcoxon

Historiquement, le test de Wilcoxon (1945) est antérieur à celui de Mann et Whitney (1947). Ils sont totalement équivalents. Pour le test de Wilcoxon, nous devons choisir un groupe de référence, par convention le 1^{er}, avec $n_1 < n_2$, et si $n_1 = n_2$, il faut que $S_1 < S_2$ (voir [6], 728). La statistique de Wilcoxon W_s correspond simplement à la somme des rangs S_1 . Sa variance est strictement identique à la variance de la statistique de Mann et Whitney $V(W_s) = V(U)$ (ou $\tilde{V}(W_s) = \tilde{V}(U)$ si nous souhaitons tenir compte des ex-aequo). Son espérance dépend du groupe de référence c.-à-d.

$$E(W_s) = \frac{n_1(n_1 + n_2 + 1)}{2} \quad (2.13)$$

Pour l'approximation normale, la statistique centrée et réduite est construite exactement de la même manière. Pour un test bilatéral, nous utilisons (si l'on ne veut pas introduire la correction de continuité) :

$$|Z| = \frac{W_s - E(W_s)}{\sqrt{V(W_s)}} \quad (2.14)$$

Nous constaterons alors que les résultats concordent parfaitement. Cela n'est guère étonnant, en effet, on montre que les statistique peuvent se déduire l'une de l'autre (voir [6], page 733) c.-à-d.

$$U = \frac{n_1(n_1 + 2n_2 + 1)}{2} - W_s \quad (2.15)$$

ou, à l'inverse (voir [11], page 343)

$$W_s = n_1 n_2 + \frac{n(n+1)}{2} - U \quad (2.16)$$

Exemple 13 (Exemple : tester le degré d'anxiété).

Reprenons notre exemple d'enfants inquiets. Puisque $n_1 = 16 < n_2 = 23$, le premier groupe "Tradition orale = absent" devient le groupe de référence. Des résultats précédents (Figure 2.5), nous pouvons extraire :

- $W_s = S_1 = 200$
- $E(W_s) = \frac{16(16+23+1)}{2} = 320$
- Par curiosité, vérifions

$$U = \frac{n_1(n_1 + 2n_2 + 1)}{2} - W_s = \frac{16(16 + 2 \times 23 + 1)}{2} - 200 = 64$$

La correspondance est vérifiée.

– En introduisant la correction pour ex-aequo, la statistique centrée et réduite s'écrit

$$|Z| = \frac{|W_s - E(W_s)|}{\sqrt{\tilde{V}(W_s)}} = \frac{|200 - 320|}{\sqrt{1209.1606}} = 3.4510$$

Soit exactement la même valeur que pour le test de Mann et Whitney.

2.1.7 Une autre vision du test de Mann et Whitney

Jusqu'à présent, nous avons montré le test de Mann et Whitney comme un test basé sur une statistique de rang linéaire, une alternative totalement équivalente au test de Wilcoxon. Cela est vrai bien entendu puisque les tests produisent exactement les mêmes résultats. En revanche, historiquement, il provient d'une autre approche. Nous nous contentons de la décrire pour l'instant. Nous aurons l'occasion de mieux exploiter cette présentation lorsque nous nous intéresserons aux tests de rangs robustes.

Supposons que le 1^{er} groupe sert de référence. Pour une observation $n^o i$ de Ω_1 , nous définissons la quantité

$$P_i = \sum_{j=1}^{n_2} I(x_{i1} > x_{j2}) \quad (2.17)$$

où $I(x_{i1} > x_{j2})$ prend la valeur 1 si la condition est vraie, 0 sinon. P_i correspond donc au nombre d'observations du second échantillon qui est plus petit que x_{i1} .

Pour chaque observation de l'échantillon de référence, nous pouvons calculer la quantité P_i . Un indicateur global du positionnement des individus du groupe de référence sera dès lors

$$S_P = \sum_{i=1}^{n_1} P_i \quad (2.18)$$

S_P correspond à la statistique U de Mann et Whitney. Il comptabilise le nombre total de couples de points (x_{i1}, x_{j2}) vérifiant l'inégalité $(x_{i1} > x_{j2})$ (voir [3], page 123-124; [11], page 343). Cette manière de calculer est totalement équivalente à l'approche basée sur la somme des rangs (section 2.1.3).

Exemple 14 (IMC selon le sport).

Un petit exemple vaut mieux qu'un long discours. Nous reprenons notre exemple des lycéens sportifs (ou non). Nous avons trié les données selon la variable d'intérêt croissante (Figure 2.6), nous avons rajouté une nouvelle colonne de manière à comptabiliser les individus de l'autre groupe portant une valeur supérieure à la valeur courante. Le groupe de référence est "Sport = DAILY" :

- La plus petite observation correspond à un individu de type "Daily", avec $x_{11} = 22.8$. Le nombre d'observations de type "Never" prenant une valeur inférieure est bien évidemment égal à $P_1 = 0$.
- La 2^{eme} observation de type "Daily" prend la valeur $x_{21} = 23.4$ (c'est la 3^{eme} observation dans l'échantillon global). Le nombre d'observations de type "Never" qui présentent une valeur inférieure est égal à $P_2 = 1$.

Autre vision de la statistique de Mann et Whitney		
IMC	Sport	DAILY Pi
22.8	DAILY	0
23	NEVER	-
23.4	DAILY	1
23.6	DAILY	1
23.7	DAILY	1
24.8	DAILY	1
26	NEVER	-
26.1	DAILY	2
26.3	NEVER	-
27.3	NEVER	-
28.7	NEVER	-
30.2	DAILY	5
33.5	NEVER	-
35.3	NEVER	-
Somme		11

Fig. 2.6. Autre manière de calculer la statistique de Mann et Whitney

– Nous continuons ainsi, pour aboutir à S_P et par conséquent à U

$$U = S_P = \sum_{i=1}^{n_1} P_i = 0 + 1 + 1 + 1 + 1 + 2 + 5 = 11$$

Ce qui correspond à la statistique calculée à partir de la somme des rangs (Figure 2.4).

2.1.8 Intérêt du test de Wilcoxon-Mann-Whitney

Le test de Wilcoxon-Mann-Whitney a un très bon comportement général. Certes, lorsque la distribution sous-jacente est donnée est gaussienne, il est un peu moins puissant que le test de Student (paramétrique). Il le surclasse en revanche dans toutes les autres configurations (voir [3], page 154).

Mais son principal avantage est ailleurs : c'est une technique de référence. Il est impossible d'y échapper dès que l'on parle de test non paramétrique de comparaison de 2 populations. Ne serait-ce que pour cette raison, il faut la connaître et bien comprendre son mode de fonctionnement.

2.1.9 Sorties des logiciels

Tanagra : Comparaison de l'anxiété

Nous reprenons notre exemple de comparaison de l'anxiété (Figure 2.5). Il faut garder en filigrane cette feuille EXCEL pour apprécier au mieux les informations produites par le logiciel. Nous souhaitons traiter les données avec le composant **MANN-WHITNEY COMPARISON** de TANAGRA. Le logiciel produit les résultats suivants (Figure 2.7). Voyons-en le détail :

- Il y a $n_1 = 16$ "tradition orale = absente" et $n_2 = 23$ "tradition orale = présente". Au total nous avons $n = 16 + 23 = 39$ observations.
- Un tableau intermédiaire résume les caractéristiques des groupes, à savoir la moyenne, la somme des rangs, et la moyenne des rangs. Ainsi, nous avons $\bar{r}_1 = 12.5$ et $\bar{r}_2 = 25.2174$.

Results								
Niveau_Anxiété	Tradition orale	Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	64.00000
		absent	16	8.9375	200.0	12.5000	E(U)	184.00000
		present	23	12.2174	580.0	25.2174	V(U)	1209.16059
		All	39	10.8718	780.0	20.0000	Z	3.45095
							P(> Z)	0.00056

Fig. 2.7. Tanagra - Test de Mann et Whitney pour les données "Anxiété"

- La statistique U de Mann-Whitney est égal à $U = 64$. Sous H_0 , son espérance est $E(U) = 184$ et sa variance $V(U) = 1209.16059$. Ainsi, nous pouvons produire la valeur absolue de la statistique centrée et réduite

$$|Z| = \frac{|64 - 184|}{\sqrt{1209.16059}} = 3.45095$$

Première information importante, Tanagra fournit directement la variance ajustée pour les ex-aequo. Seconde information importante : Tanagra n'intègre pas en revanche la correction de continuité dans le calcul de Z .

- La probabilité critique du test est $p = 0.00056$
- Si nous souhaitons introduire la correction de continuité, à partir des informations produites par le logiciel, nous n'avons qu'à former

$$|Z'| = \frac{|64 - 184| - 0.5}{\sqrt{1209.16059}} = 3.4366$$

Avec une p-value $p = 0.00029$

Tanagra, SAS et R : Comparaison des salaires masculins

Nous souhaitons maintenant comparer les salaires masculins "Salaire.Homme" selon l'octroi du prêt pour notre fichier "Crédit" (Figure 0.1). Le test de Kolmogorov-Smirnov nous a indiqué que les fonctions de répartition sont significativement différentes (section 1.3.2, 30). Nous vérifions ici, à l'aide du test de Wilcoxon-Mann-Whitney, que l'écart est imputable à une différenciation des paramètres de localisation.

TANAGRA fournit les résultats suivants (Figure 2.8), sachant que "Acceptation = non" est le groupe de référence puisqu'il est associé à l'effectif le plus faible :

Results								
Sal.Homme	Acceptation	Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	149.00000
		oui	34	7.5750	990.0	29.1176	E(U)	272.00000
		non	16	7.2281	285.0	17.8125	V(U)	2311.22286
		All	50	7.4640	1275.0	25.5000	Z	2.55849
							P(> Z)	0.01051

Fig. 2.8. Tanagra - Test de Mann et Whitney pour les données "Crédit"

- $n_1 = 16$, $S_1 = 285$ et $\bar{r}_1 = 17.8125$. De même, $n_2 = 34$, $S_2 = 990$ et $\bar{r}_2 = 29.1176$.
- Nous pouvons en déduire $U_1 = S_1 - \frac{n_1(n_1+1)}{2} = 285 - \frac{16(16+1)}{2} = 149$ et $U_2 = S_2 - \frac{n_2(n_2+1)}{2} = 990 - \frac{34(34+1)}{2} = 395$. Ainsi,

$$U = \min(U_1, U_2) = \min(149, 395) = 149$$

- Sous H_0 , nous avons l'espérance $E(U) = 272$ et $V(U) = 2311.22286$. D'où la statistique centrée et réduite, toujours sans la correction de continuité

$$|Z| = \frac{|149 - 272|}{\sqrt{2311.22286}} = 2.55849$$

- La probabilité critique est $p = 0.01051$
- Oui, à 5%, il semble que le salaire de l'homme soit différent dans les deux groupes.



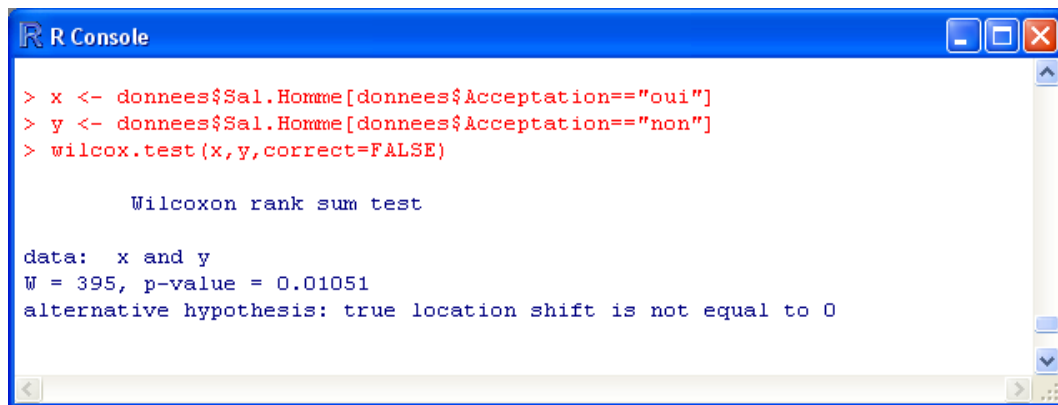
Fig. 2.9. SAS - Test de Mann et Whitney pour les données "Crédit"

SAS (version 9.1) fournit les mêmes résultats (Figure 2.9). Sauf qu'il se base sur la statistique de Wilcoxon, $W_s = S_1 = 285$. Nous avons vu qu'au final la région critique est définie exactement de la même manière que pour le test de Mann-Whitney. Par défaut, il produit la statistique centrée et réduite avec la correction de continuité. L'option "CORRECT = NO" permet de revenir à la statistique non corrigée. SAS ajoute une seconde approximation t . Il produit également le test de Kruskal-Wallis qui est une généralisation à $K \geq 2$ populations du test de Wilcoxon-Mann-Whitney.

R est plus que laconique (Figure 2.10). La statistique W est produite de la manière suivante

$$W = S_2 - \frac{n_2(n_2 + 1)}{2} = 990 - \frac{34(34 + 1)}{2} = 395$$

La probabilité critique associée est toujours $p = 0.01051$. Par défaut, R introduit aussi la correction de continuité, l'option "CORRECT=FALSE" permet de s'en affranchir.



```

R Console
> x <- donnees$Sal.Homme[donnees$Acceptation=="oui"]
> y <- donnees$Sal.Homme[donnees$Acceptation=="non"]
> wilcox.test(x,y,correct=FALSE)

      Wilcoxon rank sum test

data:  x and y
W = 395, p-value = 0.01051
alternative hypothesis: true location shift is not equal to 0

```

Fig. 2.10. R - Test de Mann et Whitney pour les données "Crédit"

2.2 Test de Fisher - Yates - Terry - Hoeffding (FYTH)

2.2.1 Principe, statistique de test

Appelé aussi test des codes normaux, le test de FYTH utilise une statistique linéaire de rang. Sans restreindre notre propos, notre groupe de référence sera le 1^{er}, avec $n_1 < n_2$, et en cas d'égalité, c'est le groupe proposant la plus petite somme des codes.

La statistique du test de FYTH s'écrit

$$C = \sum_{i=1}^{n_1} E(r_{i1}) \quad (2.19)$$

où $E(r_i)$ est l'espérance mathématique de la statistique de rang r_i dans un échantillon de taille n . Cette stratégie vise à normaliser les données après un passage par les rangs. Les codes (scores) transforment les rangs en quantiles de la loi normale.

Le calcul du code est réalisé en deux étapes :

1. Tout d'abord nous estimons la *valeur espérée* associée au rang dans la fonction de répartition. Il s'agit de la lisser en référence à la fonction de répartition sous-jacente⁶. La formule générique usuelle est (voir [1], page 373 à 375)

$$F_i = \frac{r_i + a}{n + 2a + 1}$$

Pour un échantillon prélevé dans une population normale, nous fixerons $a = -3/8$. Dans le cas du test des codes normaux, nous aurons

$$F_i = \frac{r_i - 0.375}{n + 0.25}$$

2. Dans un deuxième temps, pour obtenir les codes, nous appliquons la fonction de répartition normale inverse (centrée réduite) à la valeur F_i , soit

$$E(r_i) = \Phi^{-1}(F_i)$$

⁶ La même stratégie est utilisée pour la construction des graphiques QQ-plot.

2.2.2 Approximation normale pour les grands effectifs

L'espérance et la variance⁷ de la statistique du test s'écrivent (voir [3], page 150) :

$$E(C) = 0 \quad (2.20)$$

$$V(C) = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^n [E(r_i)]^2 \quad (2.21)$$

En cas d'ex-aequo, la formule de la variance doit être ré-écrite de la manière suivante

$$\tilde{V}(C) = \frac{n_1 n_2}{n(n-1)} \sum_{g=1}^G t_g \times [\bar{E}_g]^2 \quad (2.22)$$

Où \bar{E}_g est la moyenne des codes des observations ayant le même rang (ex-aequo).

Dès que les effectifs sont suffisamment élevés (en pratique $n_1, n_2 > 10$ ou $n > 20$), nous pouvons passer à l'approximation normale. La définition de la région critique dépend du sens du test (unilatéral à gauche ou à droite, ou bilatéral). La démarche est exactement la même que pour le test de Wilcoxon - Mann et Whitney.

Pour un test bilatéral, nous utiliserons donc la statistique centrée réduite

$$Z = \frac{C}{\sqrt{V(C)}}$$

Et la région critique au risque α s'écrit

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

Remarque 5 (Intérêt du test de Fisher-Yates-Terry-Hoeffding). Lorsque l'effectif est assez grand, le test des codes normaux est particulièrement efficace. Il surclasse le test de Wilcoxon Mann Whitney. Ce dernier garde néanmoins l'avantage sur les petits effectifs, ou lorsque l'hypothèse de normalité sous-jacente n'est absolument pas crédible. Dans le cadre gaussien, le test FYTH a une puissance similaire au test de Student de comparaison de moyennes (voir [1], page 319). On peut se demander néanmoins l'intérêt de ce test non paramétrique puisque son homologue paramétrique fonctionne parfaitement dans ce cas. L'argument en faveur est de dire "il est aussi bon que le test paramétrique lorsque les données sont gaussiennes, il est plus robuste lorsque cette hypothèse n'est plus respectée"⁸. Nous ajouterons un second atout en faveur du test de FYTH, il est peu sensible aux points aberrants. Nous évacuons ainsi un écueil fort qui peut fausser les résultats des tests de comparaisons de populations.

⁷ Attention, pour la variance, contrairement à la statistique du test, la somme est réalisée pour les n observations de l'échantillon.

⁸ Voir <http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/vanderwa.htm>; le commentaire concerne le test de Van der Waerden, mais il est transposable au test de Fisher-Yates-Terry-Hoeffding

Exemple 15 (Obésité des lapins).

Deux groupes de 10 lapins ($n_1 = n_2 = 10$, $n = 10 + 10 = 20$) ont suivi un régime alimentaire riche en cholestérol. On leur a fait alors subir 2 types de traitements (A et B), on mesure leurs caractéristiques grasses à la sortie de l'expérimentation⁹.

X	Traitement	Rang	F _i	E(r _i)	E ² (R _i)
8	A	1	0.0309	-1.87	3.49
13	A	3	0.1296	-1.13	1.27
15	A	4	0.1790	-0.92	0.84
21	A	7	0.3272	-0.45	0.20
23	A	9	0.4259	-0.19	0.03
24	A	10	0.4753	-0.06	0.00
25	A	11	0.5247	0.06	0.00
26	A	12	0.5741	0.19	0.03
28	A	14	0.6728	0.45	0.20
29	A	15	0.7222	0.59	0.35
12	B	2	0.0802	-1.40	1.97
18	B	5	0.2284	-0.74	0.55
19	B	6	0.2778	-0.59	0.35
22	B	8	0.3765	-0.31	0.10
27	B	13	0.6235	0.31	0.10
30	B	16	0.7716	0.74	0.55
31	B	17	0.8210	0.92	0.84
32	B	18	0.8704	1.13	1.27
33	B	19	0.9198	1.40	1.97
34	B	20	0.9691	1.87	3.49
			Somme	17.63	

n	20
n1	10
S1	-3.3261
n2	10
S2	3.3261
C	-3.3261
V(C)	4.6404
Z	1.5440
p-value	0.1226

Fig. 2.11. Fisher-Yates-Terry-Hoeffding - Comparaison des lapins grassouillets

Voici le détail des opérations (Figure 2.11) :

- Le groupe correspondant à "Traitement = A" nous sert de référence.
- Dans le tableau, nous distinguons les 2 sous-échantillons avec la colonne "Traitement".
- Nous attribuons les rangs bruts r_i (colonne "Rang"), nous en déduisons les fréquences espérées, soit
$$F_i = \frac{r_i - 0.375}{n + 0.25}$$
- Puis nous appliquons l'inverse de la loi normale centrée réduite $E(r_i) = \Phi^{-1}(F_i)$.
- Ainsi, pour la 1^{ère} observation, nous obtenons les valeurs : $x_1 = 8$, $r_1 = 1$, $F_1 = 0.0309$ et $E(r_1) = -1.87$; pour la 2^{ème} observation, $x_2 = 13$, $r_2 = 3$, $F_2 = 0.1296$ et $E(r_2) = -1.13$; etc.
- Le premier groupe étant la référence ($n_1 = 10$), la statistique de FYTH est obtenu avec la somme des codes limités au groupe de référence

$$C = (-1.87) + (-1.13) + \dots + 0.59 = -3.3261$$

- Pour obtenir la variance, nous calculons le carré des codes $[E(r_i)]^2$, et nous réalisons la somme sur l'ensemble des observations que nous pré-multipions par le facteur adéquat, soit

$$V(C) = \frac{10 \times 10}{20(20 - 1)} (3.49 + 1.27 + \dots + 1.97 + 3.49) = \frac{10 \times 10}{20(20 - 1)} \times 17.63 = 4.6404$$

- Pour un test bilatéral, nous calculons la valeur absolue de la statistique de test

⁹ Voir http://hdelboy.club.fr/Nonparam.htm#4-fisher_yates_terry; les tables statistiques sont disponibles dans le logiciel diffusé librement par le site.

$$Z = \frac{|-3.3261|}{\sqrt{4.6404}} = 1.5440$$

- La probabilité critique afférente est $p = 0.1226$. Au risque 5%, on ne peut pas conclure que les traitements ont une efficacité différente. A 10% non plus d'ailleurs. Ces pauvres lapins n'ont plus qu'à se mettre au footing.

2.3 Test de Van der Waerden

Principe, statistique de test et approximation normale

Le test de Van der Waerden est une version simplifiée du test de FYTH. Au lieu d'utiliser l'espérance de la statistique de rang, il propose une approche plus fruste en estimant la fréquence cumulée à l'aide de la quantité

$$F_i = \frac{r_i}{n+1}$$

Avant d'appliquer la fonction inverse la loi normale centrée réduite. Le code s'écrit donc

$$s_i = \Phi^{-1}\left(\frac{r_i}{n+1}\right)$$

Récapitulons tout cela. La statistique du test s'écrit

$$D = \sum_{i=1}^{n_1} \Phi^{-1}\left(\frac{r_{i1}}{n+1}\right) \quad (2.23)$$

Ses paramètres sont (sans ou avec prise en compte des ex-aequo) :

$$E(D) = 0 \quad (2.24)$$

$$V(D) = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^n [\Phi^{-1}\left(\frac{r_i}{n+1}\right)]^2 \quad (2.25)$$

$$\tilde{V}(D) = \frac{n_1 n_2}{n(n-1)} \sum_{g=1}^G t_g \times [\bar{s}_g]^2 \quad (2.26)$$

Où \bar{s}_g est la moyenne des codes des observations ayant la même valeur (ex-aequo).

Remarque 6 (Intérêt du test de Van der Waerden). Les tests de FYTH et Van der Waerden sont asymptotiquement équivalents. Les commentaires émis concernant les propriétés du précédent test, notamment son efficacité pour des populations gaussiennes, sont également de mise ici (voir [3], page 154, pour les comparaisons sur des données simulées.).

X	Traitement	Rang	$F_{i-r_i/(n+1)}$	$\Phi(\sqrt{12}(F_i - 0.5))$	$\Phi(\sqrt{12}(F_i - 0.5))^2$
8	A	1	0.0476	-1.67	2.78
13	A	3	0.1429	-1.07	1.14
15	A	4	0.1905	-0.88	0.77
21	A	7	0.3333	-0.43	0.19
23	A	9	0.4286	-0.18	0.03
24	A	10	0.4762	-0.06	0.00
25	A	11	0.5238	0.06	0.00
26	A	12	0.5714	0.18	0.03
28	A	14	0.6667	0.43	0.19
29	A	15	0.7143	0.57	0.32
12	B	2	0.0952	-1.31	1.71
18	B	5	0.2381	-0.71	0.51
19	B	6	0.2857	-0.57	0.32
22	B	8	0.3810	-0.30	0.09
27	B	13	0.6190	0.30	0.09
30	B	16	0.7619	0.71	0.51
31	B	17	0.8095	0.88	0.77
32	B	18	0.8571	1.07	1.14
33	B	19	0.9048	1.31	1.71
34	B	20	0.9524	1.67	2.78
			Somme	15.09	

n	20
n1	10
S1	-3.0462
n2	10
S2	3.0462
C	-3.0462
V(C)	3.9716
Z	1.5285
p-value	0.1264

Fig. 2.12. Van der Waerden - Comparaison des lapins grassouillets

Exemple 16 (Obésité des lapins).

Reprenons notre exemple du traitement de l'obésité des lapins utilisé pour illustrer le test de Fisher-Yates-Terry-Hoeffding. La structure de la feuille de calcul est exactement la même, seuls les codes sont modifiés, aboutissant à une statistique de test très légèrement différente (Figure 2.12). Voyons ce qu'il en est :

- Les valeurs X , la variable indiquant le groupe d'appartenance et les rangs bruts sont les mêmes.
- Pour la fréquence cumulée, nous utilisons maintenant

$$F_i = \frac{r_i}{n+1}$$

- Puis nous appliquons la fonction de répartition normale inverse pour obtenir les codes

$$s_i = \Phi^{-1}(F_i)$$

- La somme des codes limité au membres du premier groupe correspond à la statistique de Van der Waerden

$$D = (-1.67) + (-1.07) + \dots + 0.57 = -3.0462$$

- Pour obtenir la variance, nous formons le carré des codes s_i^2 , puis nous sommes les valeurs sur l'ensemble de l'échantillon, en pré multipliant par le facteur adéquat

$$V(D) = \frac{10 \times 10}{20(20-1)} \times (2.78 + 1.14 + \dots + 1.71 + 2.78) = \frac{10 \times 10}{20(20-1)} \times 15.09 = 3.9716$$

- Nous en déduisons

$$|Z| = \frac{|-3.0462|}{\sqrt{3.9716}} = 1.5285$$

- Avec une probabilité critique $p = 0.1264$. Au niveau de signification 5%, nous ne pouvons pas rejeter l'hypothèse nulle.
- Nous remarquerons la très forte similitude des résultats avec ceux de Fisher-Yates-Terry-Hoeffding (Figure 2.11). A la lumière des calculs réalisés, le contraire eut été très étonnant, ils diffèrent uniquement par la manière d'estimer la quantité F_i .

Exemple 17 (Anxiété des enfants).

Ce fichier, déjà étudié précédemment (section 2.1.5), est intéressant car il montre comment les ex-aequo sont traités pour le test de Van der Waerden, la démarche est identique pour le test de Fisher-Yates-Terry-Hoeffding. La principale idée à retenir est que **les scores des individus portant la même valeur doivent être identiques**. Il faut donc mettre en place une stratégie des *scores moyens*. C'est ce que nous mettrons en avant dans cet exemple. Bien sûr, il faudra également tenir compte de la correction lors du calcul de la variance.

Tradition	Anxiete	Rang brut	F_i	s_i	s'_i
absent	6	1	0.025	-1.960	-1.802
present	6	2	0.050	-1.645	-1.802
absent	7	3	0.075	-1.440	-1.168
absent	7	4	0.100	-1.282	-1.168
absent	7	5	0.125	-1.150	-1.168
absent	7	6	0.150	-1.036	-1.168
absent	7	7	0.175	-0.935	-1.168
absent	8	8	0.200	-0.842	-0.717
absent	8	9	0.225	-0.755	-0.717
present	8	10	0.250	-0.674	-0.717
present	8	11	0.275	-0.598	-0.717
absent	9	12	0.300	-0.524	-0.524
absent	10	13	0.325	-0.454	-0.256
absent	10	14	0.350	-0.385	-0.256
absent	10	15	0.375	-0.319	-0.256
absent	10	16	0.400	-0.253	-0.256
present	10	17	0.425	-0.189	-0.256
present	10	18	0.450	-0.126	-0.256
present	10	19	0.475	-0.063	-0.256
present	11	20	0.500	0.000	0.031
present	11	21	0.525	0.063	0.031
absent	12	22	0.550	0.126	0.288
absent	12	23	0.575	0.189	0.288
present	12	24	0.600	0.253	0.288
present	12	25	0.625	0.319	0.288
present	12	26	0.650	0.385	0.288
present	12	27	0.675	0.454	0.288
absent	13	28	0.700	0.524	0.638
present	13	29	0.725	0.598	0.638
present	13	30	0.750	0.674	0.638
present	13	31	0.775	0.755	0.638
present	14	32	0.800	0.842	0.938
present	14	33	0.825	0.935	0.938
present	14	34	0.850	1.036	0.938
present	15	35	0.875	1.150	1.290
present	15	36	0.900	1.282	1.290
present	15	37	0.925	1.440	1.290
present	16	38	0.950	1.645	1.645
present	17	39	0.975	1.960	1.960

g	valeur	t_g	s_barre_g	s_barre_g^2	t_g * moy^2(s_i)
1	6	2	-1.802	3.249	6.497
2	7	5	-1.168	1.365	6.827
3	8	4	-0.717	0.515	2.058
4	9	1	-0.524	0.275	0.275
5	10	7	-0.256	0.065	0.457
6	11	2	0.031	0.001	0.002
7	12	6	0.288	0.083	0.496
8	13	4	0.638	0.407	1.628
9	14	3	0.938	0.879	2.637
10	15	3	1.290	1.665	4.996
11	16	1	1.645	2.706	2.706
12	17	1	1.960	3.841	3.841
Somme					32.4210

n	39
n1 (absent)	16
S(moy(s_i))	-9.4126
n2 (present)	23
S((moy(s_i)))	9.4126
D	-9.4126
$\sqrt{V(D)}$	8.0506
Z	3.3174
p-value	0.00091

Fig. 2.13. Van der Waerden - Exemple de gestion des ex-aequo

Voyons cela dans le détail (Figure 2.13) :

- Le tableau de données est trié selon la variable d'intérêt "Anxiété".
- Nous assignons les rangs bruts, en numérotant simplement les individus dans l'ordre des lignes.
- Puis nous calculons la quantité $F_i = \frac{r_i}{n+1}$, et nous appliquons l'inverse de la fonction de répartition de la loi normale $s_i = \Phi^{-1}(F_i)$

- Arrive l'étape importante, pour que les individus présentant les mêmes valeurs se voient attribués les mêmes codes, nous effectuons une péréquation dans les sous-groupes (d'ex-aequo) en calculant la moyenne des codes. Pour les 2 premiers individus correspondants à la valeur $v_1 = 6$,

$$x_1 = x_2 = 6 \rightarrow s'_1 = s'_2 = \frac{(-1.960) + (-1.645)}{2} = -1.802$$

Pour les 5 individus suivants correspondants à la valeur $v_2 = 7$

$$x_3 = \dots = x_7 = 7 \rightarrow s'_3 = s'_4 = s'_5 = s'_6 = s'_7 = \frac{-1.440 - 1.282 - 1.150 - 1.036 - 0.935}{5} = -1.168$$

Etc.

- Ce processus est très important. D'une part, il nous assure que les individus ex-aequo portent le même score. D'autre part, le mode de calcul est indépendant de l'ordonnement des observations dans les sous-groupes. La colonne s'_i correspond aux *scores moyens* des individus.
- Il reste à produire les indicateurs récapitulatifs : $n = 39$, $n_1 = 16$, $n_2 = 23$
- La somme des scores relatifs à la classe "Tradition = absent" qui sert de référence est

$$S_1 = \sum_{i=1}^{n_1=16} s'_{i1} = -9.4126$$

Nous aurions pu faire de même pour l'autre groupe d'individus (Tradition = présent)

$$S_2 = \sum_{i=1}^{n_2=23} s'_{i2} = 9.4126$$

- Puisque $n_1 < n_2$, le premier groupe sert de référence, la statistique du test est donc

$$D = S_1 = -9.4126$$

- Pour mettre en oeuvre l'approximation normale, nous devons calculer la variance de la statistique, la variance corrigée puisqu'il y a des ex-aequo. Pour ce faire, nous créons le tableau des valeurs uniques. Il y a $G = 12$ valeurs différentes, nous les numérotions $g = 1, \dots, 12$. Nous calculons les effectifs t_g , et les moyennes des scores pour chaque sous-groupe $\bar{s}_g = \frac{\sum_{i \in g} s'_i}{t_g}$, son carré \bar{s}_g^2 et le produit $t_g \times \bar{s}_g^2$
- La somme de cette dernière colonne $\sum_g t_g \bar{s}_g^2 = 32.4210$, nous en déduisons la variance

$$\tilde{V}(D) = \frac{16 \times 23}{39(39-1)} \times 32.4210 = 8.0506$$

- Nous pouvons maintenant former la statistique centrée réduite

$$|Z| = \frac{|D|}{\sqrt{\tilde{V}(D)}} = \frac{|-9.4126|}{\sqrt{8.0506}} = 3.3174$$

- La probabilité critique est $p = 0.00091$. Au risque 5%, nous pouvons rejeter l'hypothèse nulle d'égalité du niveau d'anxiété selon leur tradition orale.

Remarque 7 (Scores moyens et non score des rangs moyens). L'erreur à ne pas commettre dans le calcul des scores moyens serait de calculer d'abord les rangs moyens sur les ex-aequo, puis de leur appliquer la fonction score $f(\cdot)$. La statistique du test est différente dans ce cas, erronée. En effet $f(\cdot)$ est non linéaire. Même si par ailleurs nous associons effectivement les mêmes scores aux ex-aequo.

2.4 Test de rang robuste

2.4.1 Le problème de Behrens-Fisher

Les test non paramétriques visent à détecter une différenciation des distributions conditionnellement aux groupes (échantillons) selon une caractéristique de localisation. Il n'en reste pas moins que l'hypothèse nulle stipule l'égalité des fonctions de répartition, sous quelque forme que ce soit. De fait, le rejet peut effectivement correspondre à un décalage entre les paramètres de localisation, mais elle peut aussi résulter d'autres formes de différenciation.

La forme des distributions doit être la même dans les sous populations pour que les tests ci-dessus fonctionnent correctement. La dispersion en est une caractéristique importante. Pour simplifier, nous parlerons de *variance*. Sous l'hypothèse nulle, nous stipulons implicitement que les variances conditionnelles sont identiques puisque les fonctions de répartition sont les mêmes. Sous H_1 , elles le sont également puisque nous cherchons uniquement à caractériser un écart entre les paramètres de localisation. Lorsque cette restriction est respectée (variances égales dans les sous-échantillons), le test de Student de comparaison de moyennes est tout à fait indiqué pour les données gaussiennes, et pour les autres (lorsque les distributions accusent une asymétrie par exemple, la même dans les sous-échantillons), les tests non paramétriques ci-dessus (Wilcoxon-Mann-Whitney entre autres) fonctionnent parfaitement¹⁰.

Lorsque nous souhaitons comparer des paramètres de localisation, sans assumer l'égalité des dispersions, les approches ci-dessus ne sont plus adaptées. Elles fournissent des résultats erronés. Le rejet de l'hypothèse nulle n'est plus imputable (seulement) à un écart de tendance centrale. Ce problème est connu sous le nom de *problème de Behrens-Fisher*. Pour les méthodes paramétriques, nous nous tournerons vers la variante d'Aspin-Welch du test de Student ; pour les méthodes non paramétriques, nous nous tournerons vers le test de rang robuste due à Fligner-Policello (1981), variante du test de Wilcoxon-Mann-Whitney (voir [3], pages 176 à 178 ; [13], page 137 à 144).

2.4.2 Test de rang robuste de Fligner-Policello

Résumons la situation, la distribution de la variable d'intérêt n'est pas gaussienne et les dispersions conditionnelles sont différentes. Dans ce contexte, le test de Fligner-Policello (1981) permet de vérifier l'égalité des paramètres de localisation dans les sous-populations. C'est une variante du test de Mann et Whitney.

¹⁰ Voir http://findarticles.com/p/articles/mi_qa3690/is_199706/ai_n8763933 pour une longue et intéressante discussion sur le sujet.

Statistique de test

Revenons sur la définition alternative de la statistique de Mann et Whitney (section 2.1.7). U correspondait à la somme du nombre d'observations inférieure à chaque individu du groupe de référence (par convention le premier, tel que $n_1 < n_2$) :

$$S_P = \sum_{i=1}^{n_1} P_i = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(x_{i1} > x_{j2}) \quad (2.27)$$

En posant

$$\bar{P} = \frac{1}{n_1} S_P \quad (2.28)$$

Nous pouvons définir un indice de variabilité de P_i

$$V_P = \sum_{i=1}^{n_1} [P_i - \bar{P}]^2 \quad (2.29)$$

De la même manière, en prenant le second groupe comme référence, nous pouvons définir tour à tour

$$S_Q = \sum_{i=1}^{n_2} Q_i = \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} I(x_{i2} > x_{j1}) \quad (2.30)$$

$$\bar{Q} = \frac{1}{n_2} S_Q \quad (2.31)$$

$$V_Q = \sum_{i=1}^{n_2} [Q_i - \bar{Q}]^2 \quad (2.32)$$

La statistique du test des rangs robuste destiné à confronter les hypothèses $H_0 : \theta = 0$ et $H_1 : \theta \neq 0$ (décalage entre les paramètres de localisation) s'écrit (voir [13], page 139)

$$U^* = \frac{S_P - S_Q}{2\sqrt{V_P + V_Q + \bar{P} \times \bar{Q}}} \quad (2.33)$$

Région critique et distribution asymptotique

U^* a été tabulée pour des petits échantillons ($n_1, n_2 \leq 12$) (voir par exemple [13], Table K, page 347). Au delà, nous pouvons utiliser l'approximation normale, sous H_0 , U^* est distribuée selon une loi normale centrée et réduite $\mathcal{N}(0; 1)$. La région critique au risque α pour un test bilatéral s'écrit tout naturellement

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

Lien entre U^* et la statistique U de Mann et Whitney

La statistique de Fligner et Policello est en réalité une variante de la statistique de Mann et Whitney, on montre en effet la relation suivante (voir [3], page 177) :

$$U^* = \sqrt{\frac{n_1}{\hat{V}(U)}} \left(\frac{U}{n_1 n_2} - \frac{1}{2} \right) \quad (2.34)$$

où $\hat{V}(U)$ est l'estimateur *naturel* de la variance de U avec

$$\hat{V}(U) = \frac{1}{n_1 n_2^2} (V_P + V_Q + \bar{P} \times \bar{Q})$$

Traitement des ex-aequo

Si les ex-aequo appartiennent au même groupe, cela n'a aucune incidence sur les calculs. En revanche, s'ils appartiennent à des groupes différents, nous devons redéfinir la fonction $I(x_{i1} > x_{j2})$ de la manière suivante

$$I(x_{i1} > x_{j2}) = \begin{cases} 0 & \text{si } x_{i1} < x_{j2} \\ 1 & \text{si } x_{i1} > x_{j2} \\ 0.5 & \text{si } x_{i1} = x_{j2} \end{cases} \quad (2.35)$$

Aucune des autres formules ne sont modifiées.

Exemple 18 (Schizophrénie et dopamine).

Des études ont montré que l'activité de la dopamine est plus importante chez les patients schizophrènes. On a donc administré un médicament destiné à bloquer son rôle à 25 patients hospitalisés. Au bout d'un certain temps, des médecins les ont classés en 2 groupes : les patients qui sont restés psychotiques $n_1 = 10$ et ceux qui ne le sont plus $n_2 = 15$. On a alors mesuré l'activité restante de la dopamine chez les patients, on cherche à vérifier si elle est différente dans les 2 groupes¹¹.

Le tableau de données a été trié selon la variable d'intérêt croissante (dopamine). La colonne suivante indique le groupe d'appartenance du sujet (Figure 2.14) :

– Nous calculons les quantités P_i pour le groupe "Évaluation = non-psychotique". Nous obtenons

$$S_P = 6 + 12 + 12 + 13 + \dots + 15 + 15 = 130$$

Nous en déduisons $\bar{P} = \frac{130}{10} = 13.0$. De même, nous calculons la somme des carrés des écarts¹², nous obtenons $V_P = 68.00$.

¹¹ ouh là là, voici un exemple bien déprimant tiré de [13], page 141 et 142, nous l'avons choisi uniquement parce qu'il nous permet de vérifier nos calculs.

¹² en utilisant la fonction SOMME.CARRES.ECARTS(...) sous EXCEL.

dopamine	evaluation	P	Q
0.0104	non-psychotique	-	0
0.0105	non-psychotique	-	0
0.0112	non-psychotique	-	0
0.0116	non-psychotique	-	0
0.0130	non-psychotique	-	0
0.0145	non-psychotique	-	0
0.0150	psychotique	6	-
0.0154	non-psychotique	-	1
0.0156	non-psychotique	-	1
0.0170	non-psychotique	-	1
0.0180	non-psychotique	-	1
0.0200	non-psychotique	-	1
0.0200	non-psychotique	-	1
0.0204	psychotique	12	-
0.0208	psychotique	12	-
0.0210	non-psychotique	-	3
0.0222	psychotique	13	-
0.0226	psychotique	13	-
0.0230	non-psychotique	-	5
0.0245	psychotique	14	-
0.0252	non-psychotique	-	6
0.0270	psychotique	15	-
0.0275	psychotique	15	-
0.0306	psychotique	15	-
0.0320	psychotique	15	-
Somme		130	20

Calcul direct	
n1	10
S_P	130
P_barre	13.0000
V_P	68.0000
n2	15
S_Q	20
Q_barre	1.3333
V_Q	49.3333
U*	4.7395
Passage par U	
U	130
V*(U)	0.0599
U*	4.7395

Fig. 2.14. Test robuste - Exemple de la dopamine chez les schizophrènes

- Nous faisons de même pour le second groupe "Évaluation = psychotique", nous obtenons : $n_2 = 15$, $S_Q = 20$, $\bar{Q} = 1.3333$ et $V_Q = 49.3333$.
- Nous en déduisons la statistique du test

$$\begin{aligned}
 U^* &= \frac{S_P - S_Q}{2\sqrt{V_P + V_Q + \bar{P} \times \bar{Q}}} \\
 &= \frac{130 - 20}{2\sqrt{68.0000 + 49.3333 + 13.0000 \times 1.3333}} \\
 &= 4.7395
 \end{aligned}$$

- Pour un risque 5%, le seuil critique sera $u_{0.975} = 1.96$, nous sommes largement dans la région critique. Manifestement, la dose de dopamine résultante, après traitement, différencie les psychotiques des autres.
- En utilisant la seconde formulation, à partir de la statistique de Mann et Whitney avec $U = S_P = 130$ (le 1^{er} groupe sert de référence puisque $n_1 = 10 < n_2 = 15$). Avec

$$\begin{aligned}
 \hat{V}(U) &= \frac{1}{n_1 n_2^2} (V_P + V_Q + \bar{P} \times \bar{Q}) \\
 &= \frac{1}{10 \times 15^2} (68.0000 + 49.3333 + 13.0000 \times 1.3333) \\
 &= 0.0599
 \end{aligned}$$

Nous pouvons calculer

$$\begin{aligned}
U^* &= \sqrt{\frac{n_1}{\hat{V}(U)}} \left(\frac{U}{n_1 n_2} - \frac{1}{2} \right) \\
&= \sqrt{\frac{10}{0.0599}} \left(\frac{130}{10 \times 15} - \frac{1}{2} \right) \\
&= 4.7395
\end{aligned}$$

Ce test présente d'excellentes qualités. Il se compare avantageusement à la fois au test d'Aspin-Welch dès que l'on s'écarte de l'hypothèse de normalité, et au test de Wilcoxon-Mann-Whitney dès que les variances conditionnelles dans les groupes deviennent inégales.

2.5 Test de la médiane

2.5.1 Principe, statistique de test et région critique (Approche A)

Malgré les apparences, le test de la médiane, attribué à Mood (1950), repose sur une statistique linéaire de rang, avec une fonction score tout à fait particulière. Nous préciserons cette idée ultérieurement. Dans cette sous section, nous nous contenterons de la présentation classique que l'on retrouve dans la grande majorité des manuels, elle n'utilise pas les rangs (voir par exemple [13], pages 124 à 128). Elle a le mérite de la simplicité.

Le test compare explicitement un indicateur de tendance centrale des 2 sous populations, en l'occurrence la médiane m . Les hypothèses pour un test bilatéral s'écrivent :

$$\begin{aligned}
H_0 : m_1 &= m_2 \\
H_1 : m_1 &\neq m_2
\end{aligned}$$

Pour le mettre en oeuvre, il suffit de comparer la proportion d'observations relatifs au premier groupe (resp. au second) en dessous de la médiane estimée sur la globalité de l'échantillon.

Précisons la démarche. Le premier groupe correspond à n_1 observations, le second à n_2 . Nous estimons la médiane \hat{m} sur les $n = n_1 + n_2$ observations, nous formons alors le tableau de contingence suivant

Effectif	Ω_1	Ω_2	Total
$X \leq \hat{m}$	a	b	a+b
$X > \hat{m}$	c	d	c+d
Total	n_1	n_2	n

Sous H_0 , la proportion des observations en dessous de la médiane dans le premier groupe $\frac{a}{n_1}$ serait identique à celle dans le second $\frac{b}{n_2}$. Plus on s'écarte de cette situation, plus nous serons amenés à rejeter l'hypothèse nulle. Pour réaliser les calculs, selon l'effectif n , nous choisirons la procédure adaptée :

1. Lorsque $n \leq 20$, nous utilisons le test exact de Fisher.

2. Lorsque $n > 20$, on peut utiliser le test du χ^2

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (2.36)$$

Elle suit une loi du $\chi^2(1)$ à 1 degré de liberté. Nous rejetons H_0 lorsque la statistique est supérieure à la valeur critique $\chi^2_{1-\alpha}(1)$, le quantile d'ordre $1 - \alpha$.

Remarque 8 (Correction de continuité pour les petits effectifs). Pour les effectifs assez faibles, si l'on souhaite quand même passer par l'approximation du χ^2 , il est conseillé d'introduire la correction de continuité¹³ pour les tables de contingence 2×2 , la statistique s'écrit alors

$$\chi^2 = \frac{n(|ad - bc| - n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

La pertinence de cette correction reste néanmoins controversée (voir [6], pages 167 et 168).

Le test de la médiane reposerait donc sur un schéma simple : comparer la proportion des observations en dessous (et au dessus) de la médiane globale dans chaque groupe. Plus on s'écarte de l'hypothèse nulle, plus la proportion sera différente. Pour quantifier l'écart, on utilise le test d'indépendance du χ^2 calculé sur le tableau de contingence croisant l'appartenance aux groupes et le positionnement par rapport à la médiane. On retrouve **cette présentation** dans de nombreux ouvrages. Et pourtant ce **n'est pas celle qui est programmée dans tous les logiciels**. Certains d'entre eux s'en tiennent au cadre *stricto sensu* des statistiques de rang linéaires, avec une fonction score particulière¹⁴. Nous développons cette idée dans la section suivante.

2.5.2 Approche par les statistiques de rang linéaires (Approche B)

Le même test de la médiane peut être présenté à travers notre canevas des statistiques de rang linéaires (voir [3], pages 123 à 125). Si Ω_1 est le groupe de référence, la statistique du test s'écrit

$$M = \sum_{i=1}^{n_1} I(r_{i1} > \frac{n+1}{2}) \quad (2.37)$$

où $f(r_{i1}) = I(r_{i1} > \frac{n+1}{2})$ est la fonction score définie de la manière suivante

$$f(r_i) = I(r_{i1} > \frac{n+1}{2}) = \begin{cases} 1 & \text{si } r_{i1} > \frac{n+1}{2} \\ 0 & \text{si } r_{i1} \leq \frac{n+1}{2} \end{cases}$$

¹³ voir R. Rakotomalala, *Etude des dépendances - Variables qualitatives. Tableau de contingence et mesures d'association*, Université Lumière Lyon 2, version 2.0, 2008 ; http://eric.univ-lyon2.fr/~ricco/cours/cours/Dependance_Variables_Qualitatives.pdf

¹⁴ Pour être précis, SPSS (version 8.0.0) utilise l'approche basée sur les tableaux de contingence ; SAS (version 9.1) utilise l'approche basée sur les fonctions scores. Il est donc tout à fait normal, pour le cas du test de la médiane, que ces logiciels ne fournissent pas les mêmes résultats.

La distribution exacte de M suit une loi hypergéométrique. Lorsque les effectifs sont assez grand, nous pouvons utiliser l'approximation normale. Pour former la statistique centrée et réduite Z , nous avons besoin de l'espérance et de la variance de M sous H_0 , elles s'écrivent

$$E(M) = \begin{cases} \frac{n_1}{2} & \text{si } n \text{ est pair} \\ \frac{n_1(n-1)}{2n} & \text{si } n \text{ est impair} \end{cases}$$

et

$$V(M) = \begin{cases} \frac{n_1 n_2}{4(n-1)} & \text{si } n \text{ est pair} \\ \frac{n_1 n_2 (n+1)}{4n^2} & \text{si } n \text{ est impair} \end{cases}$$

La statistique Z s'écrit

$$Z = \frac{M - E(M)}{\sqrt{V(M)}} \quad (2.38)$$

Pour un test bilatéral, nous rejetons H_0 si

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

On retrouve ce calcul dans certains logiciels de statistique. C'est l'approche à privilégier. En effet, elle s'inscrit dans un cadre cohérent par rapport aux autres tests présentés. Avec le recul, on se rend compte que toutes ces techniques peuvent être comprises et implémentées suivant un seul schéma directeur. Nous détaillerons cela plus loin (chapitre 5).

2.5.3 Intérêt du test de la médiane

On pense aujourd'hui que le test de la médiane est obsolète, surtout pour les petits effectifs. En effet, il n'utilise que le positionnement des points par rapport à la médiane, éludant une partie de l'information disponible, à la différence du test de Wilcoxon-Mann-Whitney par exemple, qui utilise le positionnement relatif des points à travers leurs rangs. C'est un jugement un peu hâtif qu'il faut relativiser. Les simulations montrent que le **test de la médiane est avantageux dans le contexte des distributions symétriques à queues lourdes** (voir [3], 154). En effet, dans ce type de configuration, il apparaît judicieux de ne pas donner trop d'importance aux observations situées aux extrêmes (voir [3], page 156).

De plus, lorsque les données peuvent prendre des valeurs extrêmes et sont tronquées, seul le test de la médiane est réellement efficace (voir [13], page 124). Par exemple, nous souhaitons mesurer les performances de plusieurs moteurs sur un banc de puissance, dont la capacité est limitée à 300 chevaux. Tous les moteurs qui peuvent dépasser ce seuil se verront attribuer arbitrairement cette valeur. Les mesures sont tronquées.

2.5.4 Un exemple et deux approches

Analysons le même fichier de données à l'aide des deux approches.

Exemple 19 (Faire pleuvoir les nuages (Approche A)).

On souhaite mesurer l'effet de l'ensemencement de nitrate d'argent sur des nuages. L'objectif étant de faire pleuvoir, la variable d'intérêt est la quantité de pluie récoltée (en logarithme)¹⁵. Les nuages sont subdivisés en 2 groupes, ceux qui ont été ensemençé ("oui") et ceux qui ne l'ont pas été ("non", groupe de référence). Nous utilisons tout d'abord l'approche consistant à situer les valeurs avec la médiane empirique, calculée sur la globalité de l'échantillon (Figure 2.15) :

X	ensemencement	médiane
7.092	non	sup
6.722	non	sup
5.920	non	sup
5.645	non	sup
5.772	non	sup
5.498	non	sup
5.094	non	sup
4.996	non	sup
4.554	non	inf
4.466	non	inf
4.397	non	inf
4.227	non	inf
3.857	non	inf
3.716	non	inf
3.600	non	inf
3.367	non	inf
3.353	non	inf
3.270	non	inf
3.262	non	inf
3.195	non	inf
3.077	non	inf
2.851	non	inf
2.442	non	inf
1.589	non	inf
1.589	non	inf
0.000	non	inf
7.918	oui	sup
7.437	oui	sup
7.412	oui	sup
6.886	oui	sup
6.556	oui	sup
6.193	oui	sup
6.064	oui	sup
5.811	oui	sup
5.713	oui	sup
5.616	oui	sup
5.617	oui	sup
5.541	oui	sup
5.491	oui	sup
5.302	oui	sup
5.291	oui	sup
4.864	oui	sup
4.779	oui	sup
4.773	oui	sup
4.748	oui	inf
4.526	oui	inf
3.704	oui	inf
3.487	oui	inf
3.447	oui	inf
2.862	oui	inf
2.041	oui	inf
1.411	oui	inf

Mediane	4.761		
---------	-------	--	--

Nombre de ensemencement	ensemencement ▼		
médiane ▼	non	oui	Total
inf	18	8	26
sup	8	18	26
Total	26	26	52

KHI-2	7.6923
ddl	1

p-value	0.0055
---------	--------

Fig. 2.15. Test de la médiane - Approche par les tableaux de contingence - Exemple des nuages

– Ma médiane estimée est $\hat{m} = 4.761$

¹⁵ <http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html>

- Nous créons une nouvelle colonne dans les données, elle prend la valeur "sup" (resp. "inf") si elle est supérieure à la médiane (resp. inférieure ou égale).
- Avec l'outil "Tableaux croisés dynamiques" d'Excel, nous élaborons le tableau de contingence entre la colonne "Médiane" et la colonne "Ensemencé". Nous obtenons les effectifs a , b , etc. qui vont nous permettre de calculer la statistique du test.
- Puisque ($n = 52$) est assez grand, nous utilisons la statistique du χ^2 sans la correction de continuité, nous obtenons

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} = \frac{52(18 \times 18 - 8 \times 8)^2}{26 \times 26 \times 26 \times 26} = 7.6923$$

- Le degré de liberté est égal à 1, la probabilité critique du test est $p = 0.0055$
- Au seuil de signification 5%, nous serons amenés à rejeter l'hypothèse nulle. Nous noterons néanmoins que les résultats sont un peu différents de l'approche par les statistiques de rang linéaires que nous présentons ci-dessous.

Exemple 20 (Faire pleuvoir les nuages (Approche B)).

Reprenons le même exemple, nous réalisons maintenant les calculs basés sur les statistiques de rang (Figure 2.16) :

- Les données ont été triées de manière à distinguer le groupe témoin "Ensemencé = non" du groupe expérimental "Ensemencé = oui".
- Nous disposons de $n = 52$ individus, avec $n_1 = n_2 = 26$
- Nous calculons le rang de chaque observation. Le premier individu par exemple a le rang $r_1 = 49$, le second $r_2 = 47$, etc.
- A partir des rangs, nous construisons la colonne des scores, le seuil étant

$$seuil = \frac{n + 1}{2} = \frac{52 + 1}{2} = 26.5$$

- La statistique du test correspond à la somme des scores des observations du 1^{er} groupe (dans le 1^{er} cadre du tableau, en jaune)

$$M = \sum_{i=1}^{n_1=26} f(r_{i1}) = 1 + 1 + 1 + \dots + 0 + 0 = 8$$

- ($n = 52$) est pair, nous calculons l'espérance et la variance sous H_0

$$E(M) = \frac{n_1}{2} = \frac{26}{2} = 13.0$$

$$V(M) = \frac{n_1 n_2}{4(n - 1)} = \frac{26 \times 26}{4 \times (52 - 1)} = 3.3137$$

- La statistique centrée et réduite du test est

$$|Z| = \frac{|8.0 - 13.0|}{\sqrt{3.3137}} = 2.7467$$

- Avec une probabilité critique $p = 0.0060$.

X	ensemence	rang	score
7.092	non	49	1
6.722	non	47	1
5.920	non	43	1
5.845	non	42	1
5.772	non	40	1
5.498	non	35	1
5.094	non	31	1
4.996	non	30	1
4.554	non	25	0
4.466	non	23	0
4.397	non	22	0
4.227	non	21	0
3.857	non	20	0
3.716	non	19	0
3.600	non	17	0
3.367	non	14	0
3.353	non	13	0
3.270	non	12	0
3.262	non	11	0
3.195	non	10	0
3.077	non	9	0
2.851	non	7	0
2.442	non	6	0
1.589	non	3.5	0
1.589	non	3.5	0
0.000	non	1	0
7.918	oui	52	1
7.437	oui	51	1
7.412	oui	50	1
6.886	oui	48	1
6.556	oui	46	1
6.193	oui	45	1
6.064	oui	44	1
5.811	oui	41	1
5.713	oui	39	1
5.616	oui	37	1
5.617	oui	38	1
5.541	oui	36	1
5.491	oui	34	1
5.302	oui	33	1
5.291	oui	32	1
4.864	oui	29	1
4.779	oui	28	1
4.773	oui	27	1
4.748	oui	26	0
4.526	oui	24	0
3.704	oui	18	0
3.487	oui	16	0
3.447	oui	15	0
2.862	oui	8	0
2.041	oui	5	0
1.411	oui	2	0

n	52
n1 (non)	26
n2 (oui)	26
seuil	26.5000
M	8
E(M)	13.0000
V(M)	3.3137
Z	-2.7467
p-value	0.0060

Fig. 2.16. Test de la médiane - Statistique de rang - Exemple des nuages

- Pour un niveau de signification à 5%, nous pouvons rejeter l'hypothèse nulle. Balancer de la nitrate d'argent dans les nuages les fait pleuvoir.
- A titre de comparaison, nous montrons les sorties du logiciel SAS (Figure 2.17). Les résultats correspondent parfaitement.

On notera enfin (et surtout) que ce résultat diffère de celui de l'approche basée sur le tableau de contingence (Approche A). Même si l'écart reste faible (Approche A, $\chi^2 = 7.6923$ - Approche B, $|Z|^2 = 7.5444$), il n'est pas négligeable et peut jeter le trouble chez les praticiens non avertis.

The NPARIWAY Procedure					
Median Scores (Number of Points Above Median) pour la variable x Classée par variable ensemece					
ensemece	Nb	Somme des scores	Attendue sous H0	Écart-type sous H0	Score moyen
non	26	8.0	13.0	1.820364	0.307692
oui	26	18.0	13.0	1.820364	0.692308

Les scores moyens ont été utilisés pour les liens.

Median Two-Sample Test

Statistic	8.0000
Z	-2.7467
One-Sided Pr < Z	0.0030
Two-Sided Pr > Z	0.0060

Median One-Way Analysis

Chi-Square	7.5444
DF	1
Pr > Chi-Square	0.0060

Fig. 2.17. Test de la médiane - Statistique de rang - Exemple des nuages - Sortie SAS

Tests de rang dans un modèle de localisation pour $K \geq 2$ populations

Les tests non paramétriques de comparaison de populations peuvent être étendus à K populations ($K \geq 2$)¹. Tout comme le test de Student de comparaison de moyennes peut être généralisé en analyse de variance, une comparaison simultanée de K moyennes, nous devrions pouvoir définir des sortes d'analyse de variance sur les rangs, ou plus précisément sur les scores déduits des rangs. C'est l'objet de ce chapitre.

Dans un premier temps, reconsidérons la formulation du test d'hypothèses. On cherche à savoir si les fonction de répartition conditionnelle $F_k(X)$ sont toutes identiques, l'hypothèse nulle s'écrit

$$H_0 : F_1(X) = \dots = F_k(X) = \dots = F_K(X)$$

L'hypothèse alternative est "une des distribution au moins est différente des autres". Si l'on cherche avant tout à s'intéresser au paramètre de localisation θ_k , on peut ré-écrire l'hypothèse H_0 de la manière suivante :

$$H_0 : \theta_1 = \dots = \theta_k = \dots = \theta_K$$

L'hypothèse alternative correspond bien évidemment à "un des θ_k au moins diffère des autres, ou d'un autre". On ne manquera pas de faire le rapprochement avec les hypothèses de l'analyse de variance. A la différence que θ correspond à *un indicateur de tendance centrale*, qui n'est pas forcément (surtout pas) la moyenne. On pourrait penser à la médiane (voir [13], page 206), mais en réalité nous sommes dans un cadre encore plus générique : les paramètres de localisation θ_k servent avant tout à caractériser le décalage entre les fonctions de distribution.

3.1 Test de Kruskal-Wallis

3.1.1 Principe, statistique de test et région critique

Le test de Kruskal-Wallis est la généralisation à K populations du test de la somme des rangs de Wilcoxon-Mann-Whitney bilatéral. On le considère comme l'alternative non paramétrique de l'ANOVA

¹ Pour $K = 2$, on devrait retrouver les résultats du chapitre précédent.

dès que la distribution sous-jacente des données n'est plus gaussienne. Il est extrêmement populaire (voir [1], page 321 ; [6], page 740 et 741 ; [13], pages 206 à 212 ; [3], pages 221 à 230 ; etc... et les références en ligne, telles que http://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance, on y trouve même un lien vers l'article originel de Kruskal et Wallis (1952)!).

Le rapprochement avec l'analyse de variance est justifié jusque dans la construction de la statistique de test. Soit \bar{r} la moyenne globale des rangs, et \bar{r}_k la moyenne des rangs pour les observations du groupe $n^{\circ}k$ (moyenne conditionnelle), la statistique de Kruskal-Wallis est définie de la manière suivante

$$H = \frac{12}{n(n+1)} \sum_{k=1}^K n_k (\bar{r}_k - \bar{r})^2 \quad (3.1)$$

C'est bien l'expression d'une variabilité inter-classes c.-à-d. la dispersion des moyennes conditionnelles autour de la moyenne globale. H est forcément ≥ 0 . Si l'hypothèse nulle est vérifiée, les moyennes conditionnelles des rangs sont proches de la moyenne globale (au fluctuation d'échantillonnage près), H prend une valeur proche de 0. **La région critique correspond aux grandes valeurs de H . Plus H s'écarte de 0, plus l'hypothèse alternative sera crédible.**

Or nous savons que $\bar{r} = \frac{n+1}{2}$ de manière mécanique. Il est possible de simplifier l'expression ci-dessus. On retrouve plus couramment la formule suivante dans la littérature :

$$H = \frac{12}{n(n+1)} \sum_{k=1}^K \frac{S_k^2}{n_k} - 3(n+1) \quad (3.2)$$

où S_k est la somme des rangs des individus appartenant au groupe $n^{\circ}k$.

Concernant l'efficacité du test de Kruskal-Wallis, les commentaires relatifs aux qualités du test de Wilcoxon-Mann-Whitney sont d'actualité. Il se compare très favorablement à l'ANOVA paramétrique. Dans le cas où la distribution sous-jacente des données est gaussienne, son efficacité relative asymptotique est de $\frac{3}{\pi} \approx 95.5\%$ c.-à-d. si l'hypothèse alternative est vraie, là où il faudrait 95 observations pour que l'ANOVA détecte la réponse correcte, il en faudrait 100 pour le test de Kruskal-Wallis. Dans tous les autres cas, par exemple avec des distributions conditionnelles asymétriques², il est toujours meilleur.

Remarque 9 (Équivalence avec le test de Mann et Whitney pour $K = 2$). L'équivalence avec le test de Mann et Whitney est établie par la relation entre les statistiques de test pour $K = 2$ (voir [3], page 222) :

$$H = \frac{12n_1}{n(n+1)} \left(U - \frac{n+1}{2} \right)^2$$

De plus, le carré de la statistique centrée réduite Z du test de Wilcoxon-Mann-Whitney est égal à la statistique H de Kruskal et Wallis. Ce dernier est bien une généralisation.

² Le même type d'asymétrie dans les distributions conditionnelles, sinon ses résultats seraient faussés.

Exemple 21 (Kruskal-Wallis : un exemple pour les petits échantillons - La production de capsules). Une usine veut comparer les performances de 3 réglages différents de machines outils dans la production de capsules : une réglage standard, et 2 autres types de réglages. La variable d'intérêt est le nombre de capsules produites dans une unité de temps. Cet exemple a été utilisé par Kruskal et Wallis pour illustrer leur méthode dans leur article de référence³.

Les machines à capsules						
reglage	production	rang				
standard	340	5				
standard	345	9				
standard	330	1				
standard	342	6				
standard	338	3				
modif_1	339	4				
modif_1	333	2				
modif_1	344	8				
modif_2	347	10				
modif_2	343	7				
modif_2	349	11				
modif_2	355	12				

n	12		
r_barre	6.5		

Groupe	standard	modif_1	modif_2
n_k	5	3	4
S_k	24	14	40

S^2_k	576.0	196.0	1600.0
-------	-------	-------	--------

r_barre_k	4.8	4.7	10.0
-----------	-----	-----	------

1ere formulation		2eme formulation	
A	73.5333	B	580.5333
H	5.6564	H	5.6564

Fig. 3.1. Test de Kruskal-Wallis sur les petits effectifs - Exemple des capsules

Les données et les calculs sont résumés dans une feuille EXCEL (Figure 3.1) :

- Le nombre d'observations est $n = 12$. A chaque ligne du tableau de données correspond le type de machine, la valeur prise par la variable d'intérêt et le rang associé.
- La moyenne globale des rangs est $\bar{r} = \frac{n+1}{2} = \frac{12+1}{2} = 6.5$
- Les effectifs conditionnels sont $n_1 = 5$, $n_2 = 3$ et $n_3 = 4$. Nous calculons aisément les sommes de rangs conditionnelles : $S_1 = 24$, $S_2 = 14$, $S_3 = 40$ et leurs carrés $S_1^2 = 476$, $S_2^2 = 196$, $S_3^2 = 1600$.
- Les moyennes conditionnelles sont obtenues avec $\bar{r}_k = \frac{S_k}{n_k}$, soit $\bar{r}_1 = 4.8$, $\bar{r}_2 = 4.7$, $\bar{r}_3 = 10$
- Utilisons d'abord la première formulation (équation 3.1). Nous formons

$$A = \sum_k n_k (\bar{r}_k - \bar{r})^2 = 5 \times (4.8 - 6.5)^2 + 3 \times (4.7 - 6.5)^2 + 4 \times (10 - 6.5)^2 = 73.5333$$

- Nous en déduisons la statistique du test

$$H = \frac{12}{n(n+1)} \times A = \frac{12}{12(12+1)} \times 73.5333 = 5.6564$$

- Passons à la seconde formulation (équation 3.2). Nous calculons

$$B = \sum_k \frac{S_k^2}{n_k} = \frac{576.0}{5} + \frac{196.0}{3} + \frac{1600.0}{4} = 580.5333$$

Nous en déduisons aussi

$$H = \frac{12}{n(n+1)} \times B - 3(n+1) = \frac{12}{12(12+1)} \times 580.5333 - 3(12+1) = 5.6564$$

³ W. Kruskal, W. Wallis, *Use of ranks in one-criterion variance analysis*, JASA, vol.47, n°260, pp. 583-621, Dec., 1952.

- Nous avons bien la même valeur de H . Reste à savoir si les différences entre les rangs moyens sont significatives c.-à-d. déterminer si H est situé dans la région critique du test.
- Les effectifs étant faibles, nous devons utiliser les tables de Kruskal-Wallis (voir section D.4, figure D.4). Cette dernière est organisée de manière à ce que, pour chaque combinaison de n_k , nous ayons la probabilité critique du test pour une valeur de H donnée. C'est un peu étrange, mais on le comprend dans la mesure où pour les faibles effectifs, le nombre de combinaisons possibles peut être facilement énuméré. Dans notre table, nous avons pour une combinaison de 3 groupes correspondant aux effectifs (5, 4, 3), la probabilité critique associée à la valeur $H = 5.6564$ est $p = 0.049$.
- De fait, pour un test à 5%, nous rejetons (de très peu) l'hypothèse nulle d'égalité des paramètres de localisation.

3.1.2 Distribution asymptotique

Lorsque les effectifs sont assez élevés, en pratique $n_k > 5, \forall k$, la distribution de H peut être approximée par une loi du χ^2 à $(K - 1)$ degrés de liberté lorsque H_0 est vrai.

En effet, n'oublions pas que les sommes de rangs S_k sont asymptotiquement normales (plus généralement la somme des scores). De fait, toute statistique de la forme

$$\sum_{k=1}^K \frac{[S_k - E(S_k)]^2}{V(S_k)}$$

Suit une loi du χ^2 à $K - 1$ degrés de liberté compte tenu du fait que les quantités S_k sont reliées par une relation linéaire. La région critique du test au risque α s'écrit

$$R.C. : H \geq \chi_{1-\alpha}^2(K - 1)$$

Exemple 22 (Kruskal-Wallis : distribution asymptotique - Les hot-dogs plus ou moins salés). On s'intéresse à la teneur en sel (variable d'intérêt "Sodium") de "Hot-Dogs". Il y a 3 catégories, selon la viande qu'ils contiennent : boeuf, volaille et viande (un mélange boeuf, porc et volaille)⁴. Visuellement, il semble qu'il y ait *un petit quelque chose* si l'on se réfère aux *boîtes à moustaches* conditionnelles (Figure 3.2). Il faut confirmer ou infirmer cela avec les calculs statistiques.

Les données ont été regroupées en bloc d'appartenance, nous pouvons lancer les calculs (Figure 3.3) :

- Nous disposons de $n = 54$ observations, avec les effectifs conditionnels $n_1 = 20$, $n_2 = 17$ et $n_3 = 17$. Nous réunissons les conditions nécessaires au passage à la loi asymptotique.
- Les données sont transformées en rangs, il n'y a pas d'ex-aequo dans le fichier⁵
- Nous calculons les sommes des rangs conditionnelles, nous obtenons $S_1 = 440$, $S_2 = 478$ et $S_3 = 567$. Nous les passons au carré. Tout est réuni pour calculer la statistique du test. Nous allons utiliser la seconde formulation (équation 3.2).

⁴ Voir <http://lib.stat.cmu.edu/DASL/Datafiles/Hotdogs.html>

⁵ Il y avait des eq-aequo dans le fichier originel. Nous l'avons très légèrement modifié pour ne pas avoir à gérer ce problème à ce stade de notre exposé.

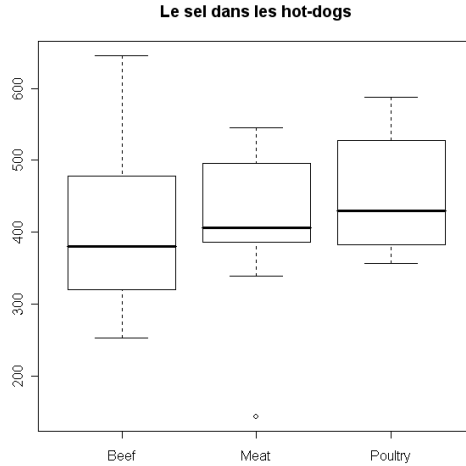


Fig. 3.2. Teneur en sel des "hot-dogs" - Boxplot

– Nous formons tout d'abord la quantité

$$B = \sum_k \frac{S_k^2}{n_k} = \frac{193600}{20} + \frac{228484}{17} + \frac{321489}{17} = 42031.3529$$

– Nous en déduisons la statistique du test

$$H = \frac{12}{n(n+1)} \times B - 3(n+1) = \frac{12}{54(54+1)} \times 42031.3529 - 3(54+1) = 4.8236$$

– Sous H_0 , H suit une loi du χ^2 à $(K-1 = 3-1 = 2)$ degrés de liberté. Pour un niveau de signification de 5%, nous devons comparer H avec le quantile $\chi_{0.95}^2(2) = 5.9915$. Les données sont compatibles avec l'hypothèse nulle d'égalité de teneur en sel des "hot-dogs".

– Résultat confirmé par la probabilité critique du test égale à $p = 0.0897$

3.1.3 Traitement des ex-aequo

Lorsque les données comportent des ex-aequo, nous utilisons le principe des rangs moyens et la statistique du test devra être corrigée. Soit G le nombre de valeurs distinctes dans le fichier ($G \leq n$). Pour la valeur $n^o g$, nous observons t_g valeurs. La statistique ajustée s'écrit

$$\tilde{H} = \frac{H}{1 - \frac{\sum_{g=1}^G (t_g^3 - t_g)}{n^3 - n}} \quad (3.3)$$

Attention, la statistique H est calculée sur les rangs modifiés (c.-à-d. avec les rangs moyens lorsqu'il y a des ex-aequo).

Exemple 23 (Un second exemple : poids des animaux à la naissance). Cet exemple est encore une fois repris de l'article fondateur des auteurs de la méthode (pages 588 et 589), ils semblent l'avoir repris

type	sodium	rang
Beef	253	2
Beef	298	3
Beef	300	4
Beef	317	5
Beef	319	6
Beef	322	7
Beef	324	8
Beef	330	9
Beef	370	15
Beef	375	17
Beef	385	20
Beef	401	26
Beef	425	29
Beef	440	33
Beef	477	36
Beef	479	37
Beef	482	38
Beef	495	39
Beef	587	52
Beef	645	54
Meat	144	1
Meat	339	10
Meat	360	14
Meat	372	16
Meat	386	21
Meat	387	22
Meat	393	24
Meat	405	27
Meat	406	28
Meat	428	31
Meat	458	34
Meat	473	35
Meat	496	40
Meat	506	41
Meat	507	42
Meat	511	43
Meat	545	49
Poultry	357	11
Poultry	358	12
Poultry	359	13
Poultry	376	18
Poultry	383	19
Poultry	388	23
Poultry	396	25
Poultry	426	30
Poultry	430	32
Poultry	513	44
Poultry	515	45
Poultry	522	46
Poultry	528	47
Poultry	542	48
Poultry	546	50
Poultry	581	51
Poultry	588	53

n	54
r_barre	27.5

	Beef	Meat	Poultry
n_k	20	17	17
S_k	440	478	567
S^2_k	193600	228484	321489

B	42031.3529
H	4.8236
ddl	2
KHI-2 0.95(2)	5.9915
p-value	0.0897

Fig. 3.3. Test de Kruskal-Wallis sur les grands effectifs - Teneur en sel des "hot-dogs"

eux-même d'un ouvrage de Snedecor⁶ Il s'agit de comparer le poids à la naissance de $K = 8$ portée de porcs. Ici également, l'intérêt est de pouvoir calibrer nos calculs⁷.

Les données ont été triées selon la variable d'intérêt croissante "poids". Nous disposons de $n = 56$ observations et $K = 8$ groupes ("portée") (Figure 3.4) :

- Nous avons comptabilisé le nombre d'observations n_k dans chaque groupe, nous obtenons $n_1 = 10$, $n_2 = 8$, etc.

⁶ G. Snedecor, *Statistical Methods*, Iowa State College Press, 1937.

⁷ Eux à l'époque ne disposaient ni d'un ordinateur, ni d'un tableur, même pas d'une calculatrice. Ils faisaient tous les calculs à la main ? avec une règle à calcul ??? C'est quand même positivement impressionnant.

poids	portée	rang
1.1	c1	1
1.2	c7	2.5
1.2	c7	2.5
1.4	c8	4
1.6	c2	5
1.9	c1	6
2.0	c1	8.5
2.0	c2	8.5
2.0	c5	8.5
2.0	c5	8.5
2.1	c5	11
2.2	c7	12.5
2.2	c7	12.5
2.3	c2	14
2.4	c2	15.5
2.4	c8	15.5
2.5	c4	18.5
2.5	c6	18.5
2.5	c7	18.5
2.5	c8	18.5
2.6	c3	23
2.6	c4	23
2.6	c5	23
2.6	c5	23
2.6	c7	23
2.8	c1	27.5
2.8	c1	27.5
2.8	c2	27.5
2.8	c4	27.5
2.9	c3	31.5
2.9	c4	31.5
2.9	c5	31.5
2.9	c6	31.5
3.0	c8	34
3.1	c3	36
3.1	c6	36
3.1	c6	36
3.2	c1	41
3.2	c2	41
3.2	c3	41
3.2	c3	41
3.2	c3	41
3.2	c4	41
3.2	c4	41
3.3	c1	47.5
3.3	c1	47.5
3.3	c3	47.5
3.3	c3	47.5
3.3	c4	47.5
3.3	c4	47.5
3.4	c3	51
3.5	c2	52.5
3.5	c2	52.5
3.6	c1	54.5
3.6	c3	54.5
4.4	c1	56

n	56
---	----

portée	c1	c2	c3	c4	c5	c6	c7	c8
n _k	10	8	10	8	6	4	6	4
S _k	317	216.5	414	277.5	105.5	122	71.5	72

S ² _k	100489	46872	171396	77006	11130	14884	5112.3	5184
S ² _{k/n_k}	10048.9	5859	17139.6	9625.8	1855	3721	852.04	1296

B	50397.396
---	-----------

H	18.464
---	--------

Comptage des valeurs distinctes			
g	valeur (v _g)	t _g	t _g ³ -t _g
1	1.1	1	0
2	1.2	2	6
3	1.4	1	0
4	1.6	1	0
5	1.9	1	0
6	2	4	60
7	2.1	1	0
8	2.2	2	6
9	2.3	1	0
10	2.4	2	6
11	2.5	4	60
12	2.6	5	120
13	2.8	4	60
14	2.9	4	60
15	3	1	0
16	3.1	3	24
17	3.2	7	336
18	3.3	6	210
19	3.4	1	0
20	3.5	2	6
21	3.6	2	6
22	4.4	1	0
Somme			960

Ajustement (C)	0.9945
----------------	--------

~H	18.565
----	--------

ddl	7
-----	---

KH12 0.95(7)	14.067
--------------	--------

p-value	0.010
---------	-------

Fig. 3.4. Test de Kruskal-Wallis avec les ex-aequo - Poids à la naissance de portées de porcs

- Puis nous avons attribué les rangs aux individus. Ils ont été corrigés selon la méthode des rangs moyens. Par exemple, les 2 individus correspondant à la valeur $v_g = 1.2$ ont récupéré le rang $\frac{2+3}{2} = 2.5$, etc.
- Nous réalisons la somme des rangs S_k par groupe, nous obtenons $S_1 = 317$, $S_2 = 216.5$, etc. Nous montons ces valeurs au carré pour disposer des S_k^2
- Nous pouvons dès lors former la somme

$$B = \sum_k \frac{S_k^2}{n_k} = 50397.396$$

– et la statistique non ajustée de Kruskal-Wallis

$$H = \frac{12}{n(n+1)} \times B - 3(n+1) = \frac{12}{56(56+1)} \times 50397.396 - 3(56+1) = 18.464$$

- Penchons nous maintenant sur l'ajustement. Il y a $G = 22$ valeurs distinctes dans le fichier. Dans le tableau "Comptage des valeurs distinctes", nous disposons des effectifs associés aux valeurs t_g , nous formons la quantité $t_g^3 - t_g$, puis la somme $\sum_g (t_g^3 - t_g) = 960$
- Le facteur d'ajustement est égal à

$$C = 1 - \frac{\sum_{g=1}^G (t_g^3 - t_g)}{n^3 - n} = 1 - \frac{960}{56^3 - 56} = 0.9945$$

Il est faible. Il le sera d'autant plus que l'effectif est élevé et que le nombre d'ex-aequo est limité. A l'extrême, s'il n'y a pas d'ex-aequo dans le fichier, nous aurons $C = 1$.

- Nous pouvons alors produire la statistique ajustée

$$\tilde{H} = \frac{H}{C} = \frac{18.464}{0.9945} = 18.565$$

- Sous H_0 , il suit asymptotiquement une loi du $\chi^2(7)$ à $ddl = 7$ degrés de liberté. Le seuil critique au risque $\alpha = 5\%$ est le quantile $\chi_{0.95}^2(7) = 14.067$. Nous rejetons l'hypothèse nulle d'égalité des poids à la naissance dans les portées.
- La probabilité critique $p = 0.01$ est cohérente avec cette conclusion.

3.1.4 Sorties des logiciels

Pour illustrer les sorties des logiciels, nous utilisons le fichier "Crédit" (Figure 0.1). Nous souhaitons savoir si le revenu par tête des ménages ("Rev.Tete") est différent selon le type de garantie supplémentaire qu'ils ont contracté ("Garantie.Supp", avec 3 modalités possibles).

TANAGRA produit les résultats suivants (Figure 3.5) :

Results								
Attribute_Y	Attribute_X	Description					Statistical test	
		Value	Examples	Average	Rank sum	Rank mean	Statistics	Value
Rev.Tete	Garantie.Supp	hypothèque	29	6.8566	688.5	23.7414	Kruskal-Wallis	1.209128
		caution	5	7.2180	152.5	30.5000	KW (corr.ties)	1.209418
		non	16	7.1062	434.0	27.1250		0.546233
		All	50	6.9726	1275.0	25.5000		

Fig. 3.5. Tanagra - Test de Kruskal-Wallis sur le fichier "Crédit"

- Pour chaque modalité de la variable indépendante "Garantie.Supp", nous disposons des effectifs, de la moyenne empirique, de la somme des rangs et de la moyenne des rangs. Ainsi, pour la première modalité "Garantie.Supp = hypothèque", nous obtenons $n_1 = 29$, $\bar{x}_1 = 6.8566$, $S_1 = 688.5$ et $\bar{r}_1 = 23.7414$.

- La statistique de Kruskal-Wallis non corrigée des ex-aequo est fournie, $H = 1.209128$ avec une probabilité critique $p = 0.546313$.
- Plus intéressante pour nous est la statistique corrigée des ex-aequo, nous avons $\tilde{H} = 1.209418$ avec $p = 0.546233$. Comme nous l'avons signalée dans l'exposé de la méthode, la correction est faible généralement.
- Il apparaît en tous les cas que l'hypothèse nulle d'égalité des revenus par tête selon la garantie supplémentaire contractée n'est pas démentie par les données.

SAS, outre les indicateurs par groupe (effectif, somme des rangs, etc.), fournit directement la statistique corrigée et la probabilité critique (Figure 3.6). **R** en fait de même (Figure 3.7).

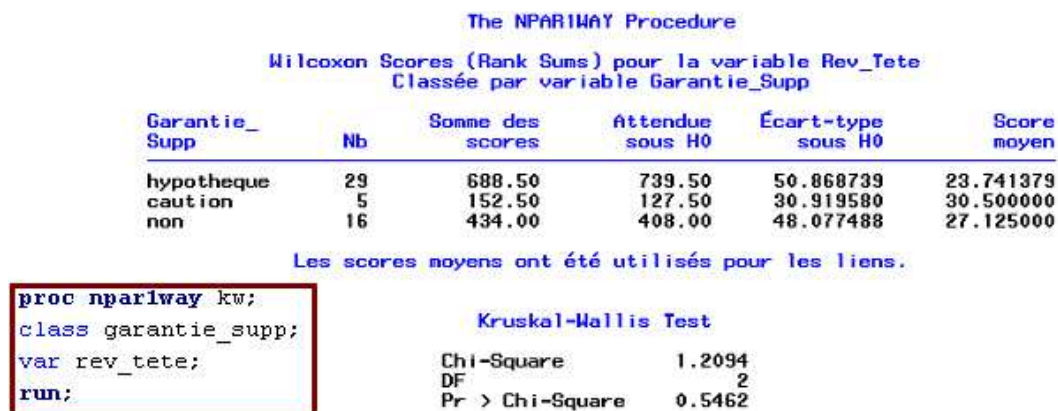


Fig. 3.6. SAS - Test de Kruskal-Wallis sur le fichier "Crédit"

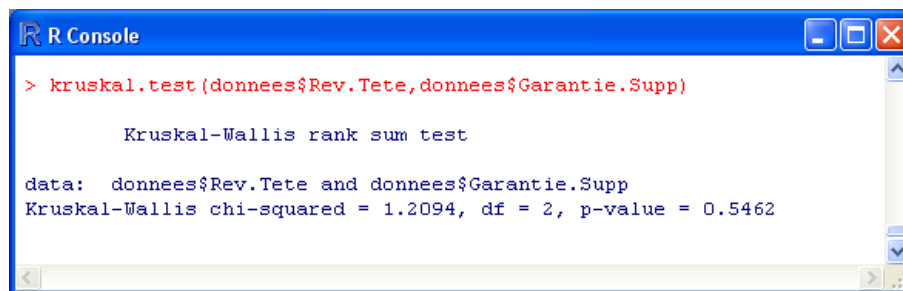


Fig. 3.7. R - Test de Kruskal-Wallis sur le fichier "Crédit"

3.2 Détermination de la source des écarts

Lorsque le test de Kruskal-Wallis aboutit au rejet de l'hypothèse nulle, cela veut dire qu'un des paramètres de localisation *au moins* est différent d'un autre. Mais nous ne disposons pas d'indications ni sur le nombre, ni sur l'identité des groupes différents. Dans cette section, nous étudions deux points de vue sur la comparaison des groupes : le premier consiste à rechercher tous les couples de groupes différents ; le second consiste à comparer chaque groupe à une référence.

3.2.1 Comparaisons multiples

La comparaison multiple consiste à comparer les paramètres de localisation θ_k de tous les groupes d'observations Ω_k . Le test s'écrit

$$H_0 : \theta_j = \theta_l$$

$$H_1 : \theta_j \neq \theta_l$$

Nous avons $\frac{K(K-1)}{2}$ comparaisons à réaliser. L'idée importante à retenir est que nous souhaitons conserver un niveau de test global compatible avec celui qui a été utilisé pour le test de Kruskal-Wallis. Il nous faut donc réduire d'autant le niveau de risque dans les comparaisons individuelles.

Si la région critique du test de Kruskal-Wallis a été définie au risque α , nous pouvons décider que les paramètres de localisation de 2 groupes sont différents si

$$|\bar{r}_j - \bar{r}_l| \geq u_{1-a} \times \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_j} + \frac{1}{n_l} \right)} \quad (3.4)$$

où u_{1-a} est le quantile d'ordre $(1-a)$ de la loi normale centrée réduite, avec

$$a = \frac{\alpha}{K(K-1)}$$

Exemple 24 (Le sel dans les hot-dogs). Reprenons notre exemple des "Hot-dogs" (Figure 3.3; le tableau de données y est, entre-autres, disponible). La probabilité critique du test est $p = 0.0897$. Pour un test à 10%, la teneur en sel des sandwiches selon la viande utilisée serait donc différente. Nous ne savons pas en revanche sur quels types de sandwiches portent les différences. C'est ce que nous souhaitons analyser maintenant (Figure 3.8) :

- Pour rappel, nous disposons de $n = 54$ observations subdivisés en $K = 3$ groupes, avec $n_1 = 20$, $n_2 = 17$ et $n_3 = 17$.
- Après le calcul de les sommes des rangs conditionnelles, nous obtenons les moyennes de rangs conditionnelles avec $\bar{r}_1 = 22.00$, $\bar{r}_2 = 28.12$ et $\bar{r}_3 = 33.35$
- Le test de Kruskal-Wallis est significatif à $\alpha = 0.1$, pour obtenir le risque pour les comparaisons individuelles, nous calculons

$$a = \frac{\alpha}{K(K-1)} = \frac{0.1}{3(3-1)} = 0.0167$$

- Le quantile d'ordre $1 - a = 1 - 0.0167$ de la loi normale centrée réduite est $u_{1-0.0167} = 2.1280$
- Pour identifier les écarts significatifs, nous réalisons les calculs de l'équation 3.4. Détaillons les pour l'opposition "Beef" vs. "Meat".
- Nous formons tout d'abord la valeur absolue de l'écart entre les rangs moyens, soit $|\bar{r}_1 - \bar{r}_2| = |22.00 - 28.12| = 6.1176$

n	54			
K	3			
	Beef	Meat	Poultry	
n_k	20	17	17	
S_k	440	478	567	
r_barre_k	22.00	28.12	33.35	
alpha	0.10			
a	0.0167			
u_a	2.1280			
Type 1	Type 2	Ecart	Seuil	Significatif
Beef	Meat	6.1176	11.0441	non
Beef	Poultry	11.3529	11.0441	oui
Meat	Poultry	5.2353	11.4831	non

Fig. 3.8. Comparaisons multiples - Teneur en sel des "Hot-Dogs"

– Nous devons la comparer avec le seuil défini par

$$u_{1-a} \times \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_j} + \frac{1}{n_l} \right)} = 2.1280 \times \sqrt{\frac{54(54+1)}{12} \left(\frac{1}{20} + \frac{1}{17} \right)} = 11.0441$$

- Manifestement, $|\bar{r}_1 - \bar{r}_2| = 6.1176 < 11.0441$. Nous ne pouvons pas affirmer que la teneur en sel d'un hot-dog au boeuf soit différent de celui d'un hot-dog à la viande (qui est un mélange improbable de différentes viandes).
- En complétant les calculs, nous nous rendons compte que le seul écart significatif survient lors de l'opposition entre "Beef" et "Poultry". On farcit les sandwiches à la volaille de sel pour que ça aille autant de goût que ceux au boeuf ? Nous laisserons les experts culinaires répondre à cette question.

Remarque 10 (D'autres formulations des comparaisons multiples). Notre présentation est dérivée du test de Kruskal-Wallis, elle est valable uniquement si les effectifs sont *suffisamment* élevés. Dans [3] (pages 236 à 240), les formulations exactes sont décrites. Mais surtout des solutions alternatives sont présentées. Nous y apprenons notamment qu'il est possible d'approximer la quantité u_{1-a} par le seuil critique du test de Kruskal-Wallis au risque α . Nous pouvons également déduire les comparaisons du test de Wilcoxon. L'avantage de cette stratégie est de permettre la construction d'intervalles de confiance des écarts entre les paramètres de localisation.

3.2.2 Comparaisons à une référence

Dans certaines situations, un des groupes Ω_c sert de référence. On cherche alors à savoir quels sont ceux qui s'en démarquent parmi Ω_k , $k \neq c$. Dans un schéma d'expérimentation, les observations Ω_c peut correspondre à un traitement standard (ou traitement de contrôle), celle utilisée jusqu'à présent. L'idée serait alors de vérifier quels sont les autres traitements qui permet d'améliorer les performances c.-à-d. augmenter significativement le rang moyen associé.

Nous avons plusieurs comparaisons à réaliser, $(K - 1)$ exactement. Il faut que le risque global α du test de Kruskal-Wallis soit maintenu, nous devons dès lors réduire le risque dans les tests individuels.

Le test d'hypothèse s'écrit

$$\begin{aligned} H_0 : \theta_k &= \theta_c \\ H_1 : \theta_k &\neq \theta_c \end{aligned}$$

La région de rejet de l'hypothèse nulle correspond à

$$|\bar{r}_k - \bar{r}_c| \geq u_{1-a} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_k} + \frac{1}{n_c} \right)} \quad (3.5)$$

où

$$a = \frac{\alpha}{2(K-1)}$$

On ne manquera pas de faire le rapprochement avec la région critique des tests dans les comparaisons multiples (équation 3.4). Finalement, la seule modification porte sur le nombre de comparaisons qui est *seulement* de $(K-1)$ ici. La définition du quantile de la loi normale utilisée dans l'équation 3.5 est ajustée en conséquence.

Remarque 11 (Test unilatéral). Nous pouvons bien entendu spécifier un test unilatéral avec $H_1 : \theta_k > \theta_r$ ou $H_1 : \theta_k < \theta_r$. Dans ce cas, la région critique sera définie de la manière suivante

$$\bar{r}_k - \bar{r}_c \gtrless u_{1-a} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_k} + \frac{1}{n_c} \right)}$$

où $a = \frac{\alpha}{K-1}$

Exemple 25 (Machines à capsules). Nous reprenons l'exemple des machines à capsules de Kruskal-Wallis (figure 3.1). Malgré que les effectifs soient relativement faibles (il existe des groupes avec moins de 5 observations), nous utilisons ces données car la configuration correspond exactement au schéma de comparaison à un traitement de contrôle. La méthode standard de production des capsules a un niveau de performances donné, on cherche à savoir si des modifications des machines va améliorer la quantité produite. L'hypothèse alternative du test est de la forme $H_1 : \theta_k > \theta_c$.

Détaillons les calculs (Figure 3.9) :

- Rappelons que le test global de Kruskal-Wallis était significatif à $\alpha = 5\%$.
- Dans notre feuille, nous calculons les effectifs et les sommes de rangs conditionnels. Nous obtenons alors les rangs moyens conditionnels $\bar{r}_1 = 4.8$, $\bar{r}_2 = 4.7$, $\bar{r}_3 = 10.0$. Le 1^{er} groupe sert de référence, $\bar{r}_c = \bar{r}_1 = 4.8$. Voyons comment s'en démarquent les autres.

Les machines à capsules					
			n	12	
			K	3	
			Groupe	standard	modif 1
			n_k	5	3
			S_k	24	14
			r_barre_k	4.8	4.7
			alpha	0.05	
			a	0.025	
			u	1.9600	
			Groupe	Ecart (standard)	Seuil
			modif_1	-0.1	5.1608
			modif2	5.2	4.7405

Fig. 3.9. Comparaisons à un traitement de contrôle - Production de capsules

- L'écart $\bar{r}_2 - \bar{r}_c = -0.1 < 0$. Nous sommes dans un schéma unilatéral, ce type de situation peut très bien survenir. Il est évident qu'il n'est pas nécessaire de réaliser le test individuel (nous l'avons quand même fait dans la feuille Excel, mais nous ne décrivons pas les résultats ici).
- Intéressons nous plutôt à l'écart entre $\bar{r}_3 - \bar{r}_c = 5.2 > 0$. Est-il significatif?
- Calculons tout d'abord le risque individuel, $a = \frac{\alpha}{K-1} = \frac{0.05}{2} = 0.025$. Nous en déduisons le quantile $u_{1-a} = u_{1-0.025} = 1.9600$.
- nous formons alors le seuil critique

$$u_{1-a} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_k} + \frac{1}{n_c} \right)} = 1.96 \times \sqrt{\frac{12(12+1)}{12} \left(\frac{1}{4} + \frac{1}{5} \right)} = 4.7405$$

Comme $\bar{r}_3 - \bar{r}_c = 5.2 > 4.7405$, nous pouvons en déduire que la modification "Modif 2" des machines outils améliore significativement la quantité de capsules produites.

3.3 Autres tests pour $K \geq 2$

3.3.1 Autant de tests que de fonctions score

Le test de Kruskal-Wallis est une généralisation du test de Wilcoxon-Mann-Whitney. En réalité, nous pouvons définir autant de tests que de fonctions score. C'est la principale information qu'il faut retenir pour le cas $K \geq 2$.

Tous les tests peuvent être généralisés en s'appuyant sur l'analogie avec l'analyse de variance. Au lieu de procéder à une ANOVA sur les rangs, nous réaliserons une ANOVA sur les scores (qui sont fonction des rangs, ne l'oublions pas), avec une démarche strictement identique. En creusant un peu plus, on se rend compte même qu'il y a un schéma directeur générique pour définir les tests à partir des scores déduits des rangs. Nous y reviendrons en détail dans le chapitre 5.

Ainsi, nous pourrions facilement généraliser les tests de Fisher-Yates-Terry-Hoeffding ou de Van der Waerden pour la comparaison de $K \geq 2$ populations. Pour les petits effectifs, nous devons nous servir des valeurs critiques dans les tables statistiques (hélas, ces statistiques étant peu usuelles, les tables sont

très rarement disponibles); pour les effectifs suffisamment grands, nous passerons par la distribution asymptotique : si H_0 est vrai, la statistique suit une loi du χ^2 à $(K - 1)$ degrés de liberté.

Prudence cependant, il semble que pour certains tests, ceux dérivés des codes normaux par exemple, cette approximation par la loi du χ^2 est de mauvaise qualité pour les effectifs modérés (voir [3], page 228). Il faut passer par d'autres lois asymptotiques.

3.3.2 Le test de la médiane généralisée

Nous consacrons une section particulière au test de la médiane car, exactement comme dans la comparaison de $K = 2$ populations, la généralisation à $K \geq 2$ décrite dans la majorité des références⁸, et implémentée dans certains logiciels⁹, n'utilise pas les rangs; la méthode implémentée dans d'autres logiciels¹⁰ s'en tiennent strictement à l'approche fondée sur les statistiques de rang linéaires. Il ne faut pas s'étonner dès lors que ces outils fournissent des résultats différents sur les données.

Nous décrivons dans cette section l'approche basée sur les tableaux de contingence, généralisation à $K \geq 2$ de la méthode décrite plus haut (Approche A - Section 2.5). Celle basée sur les rangs s'inscrit dans le cadre de l'analyse de variance sur les scores que nous détaillerons plus loin (chapitre 5).

La méthode est relativement simple :

1. Nous calculons la médiane globale \hat{m} de l'échantillon.
2. Nous formons le tableau de contingence comptabilisant les effectifs, avec en ligne, être en-deçà ou au-delà de la médiane, en colonne les catégories définissant les sous-populations. La forme générique du tableau est la suivante :

Effectif	Ω_1	\dots	Ω_k	\dots	Ω_K	Total
$X \leq \hat{m}$	o_{11}	\dots	o_{1k}	\dots	o_{1K}	$o_{1.}$
$X > \hat{m}$	o_{21}	\dots	o_{2k}	\dots	o_{2K}	$o_{2.}$
Total	$o_{.1}$	\dots	$o_{.k}$	\dots	$o_{.K}$	$o_{..} = n$

3. Sous l'hypothèse nulle, la proportion des observations situées au-dessus (ou en dessous, qu'importe) de la médiane est la même dans tous les groupes. On utilise le test d'indépendance du χ^2 pour vérifier cette assertion.
4. Pour construire cette statistique, on calcule tout d'abord l'effectif théorique dans chaque cellule du tableau sous l'hypothèse H_0 , il est égal à

$$e_{lk} = \frac{o_{l.} \times o_{.k}}{n}$$

5. La statistique s'écrit alors

$$\chi^2 = \sum_{l=1}^2 \sum_{k=1}^K \frac{(o_{lk} - e_{lk})^2}{e_{lk}} \quad (3.6)$$

⁸ voir [13], pages 200 à 206 ; http://en.wikipedia.org/wiki/Median_test

⁹ SPSS

¹⁰ <http://v8doc.sas.com/sashtml/>

6. Sous H_0 , elle suit une loi du $\chi^2(K-1)$ à $(K-1)$ degrés de liberté.
7. La région critique du test au risque α correspond aux grandes valeurs de la statistique, soit

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

Exemple 26 (Test de la médiane - Les hot-dogs plus ou moins salés).

type	sodium	mediane
Beef	253	inf
Beef	298	inf
Beef	300	inf
Beef	317	inf
Beef	319	inf
Beef	322	inf
Beef	324	inf
Beef	330	inf
Beef	370	inf
Beef	375	inf
Beef	385	inf
Beef	401	inf
Beef	425	sup
Beef	440	sup
Beef	477	sup
Beef	479	sup
Beef	482	sup
Beef	495	sup
Beef	587	sup
Beef	645	sup
Meat	144	inf
Meat	339	inf
Meat	360	inf
Meat	372	inf
Meat	386	inf
Meat	387	inf
Meat	393	inf
Meat	405	inf
Meat	406	sup
Meat	428	sup
Meat	458	sup
Meat	473	sup
Meat	496	sup
Meat	506	sup
Meat	507	sup
Meat	511	sup
Meat	545	sup
Poultry	357	inf
Poultry	358	inf
Poultry	359	inf
Poultry	376	inf
Poultry	383	inf
Poultry	388	inf
Poultry	396	inf
Poultry	426	sup
Poultry	430	sup
Poultry	513	sup
Poultry	515	sup
Poultry	522	sup
Poultry	528	sup
Poultry	542	sup
Poultry	546	sup
Poultry	581	sup
Poultry	588	sup

mediane	405.5
---------	-------

Tableau de comptage				
mediane	Beef	Meat	Poultry	Total
<= m	12	8	7	27
> m	8	9	10	27
Total	20	17	17	54

Tableau sous H ₀			
mediane	Beef	Meat	Poultry
<= m	10	8.5	8.5
> m	10	8.5	8.5

Ecart ²			
mediane	Beef	Meat	Poultry
<= m	0.40	0.03	0.26
> m	0.40	0.03	0.26

KHI-2	1.3882
-------	--------

ddl	2
-----	---

Seuil critique	5.9915
----------------	--------

p-value	0.4995
---------	--------

Fig. 3.10. Test de la médiane - Les hot-dogs plus ou moins salés

Nous reprenons notre exemple des "hot-dogs" plus ou moins salés selon la viande utilisée pour sa confection ($K = 3$ catégories de viandes). Nous disposons de $n = 54$ observations. La médiane estimée de la variable d'intérêt "sodium" est égale à $\hat{m} = 405.5$. Nous pouvons comptabiliser le positionnement des observations selon les groupes (Figure 3.10) :

- Dans le "Tableau de comptage", nous y retrouvons bien nos $n = 54$ objets, 27 d'entre eux sont en dessous de la médiane, il en est de même au dessus. C'est un résultat attendu, il doit y avoir autant d'individus en dessous et au dessus de la médiane.
- Il y a $n_1 = 20$ sandwiches composés de "Beef", 12 d'entre eux ont une teneur en sel en dessous de la médiane globale, soit $\frac{12}{20} = 60\%$. Cette proportion est de $\frac{8}{17} = 47\%$ dans le groupe "Meat"; $\frac{7}{17} = 41\%$ dans le groupe "Poultry". Ce sont ces proportions qu'il faudra comparer.
- Le "tableau sous l'hypothèse d'indépendance" est construit à partir des marges du tableau de comptage.
- Nous pouvons alors former le tableau des écarts au carré qui correspond à la formule

$$\frac{(o_{lk} - e_{lk})^2}{e_{lk}}$$

Et former la somme pour obtenir la statistique du χ^2

$$\chi^2 = 0.40 + 0.03 + 0.26 + 0.40 + 0.03 + 0.26 = 1.3882$$

- Le degré de liberté est égal à $(K - 1 = 3 - 1 = 2)$. Au risque $\alpha = 5\%$, le seuil critique du test sera $\chi_{0.95}^2(2) = 5.9915$.
- Nous sommes dans la région d'acceptation de H_0 . Les données sont compatibles avec l'hypothèse de teneur en sel identique des "hot-dogs".
- La probabilité critique est bien évidemment cohérente avec la conclusion $p = 0.4995$.

The NPARIWAY Procedure

Median Scores (Number of Points Above Median) pour la variable sodium
Classée par variable type

type	Nb	Somme des scores	Attendue sous H0	Écart-type sous H0	Score moyen
Beef	20	8.0	10.00	1.790363	0.400000
Meat	17	9.0	8.50	1.722493	0.529412
Poultry	17	10.0	8.50	1.722493	0.588235

Median One-Way Analysis

Chi-Square	1.3625
DF	2
Pr > Chi-Square	0.5060

Fig. 3.11. Sortie SAS pour le test de la médiane - Teneur en sel des "hot-dogs"

A titre de comparaison, le logiciel SAS, basé strictement sur les statistique de rangs linéaire nous fournit un $\chi^2 = 1.3625$, avec une probabilité critique égale à $p = 0.5060$ (Figure 3.11). Encore une fois, comme dans le cas $K = 2$, les résultats de l'approche par les tableaux de contingence diffèrent de ceux de l'approche basé strictement sur les statistiques de rang linéaires. L'écart n'est pas dramatique, mais il est réel. Il faut en comprendre la source pour ne pas être pris de court lorsque nous avons à analyser les sorties des logiciels.

Enfin, par rapport au test de Kruskal-Wallis, nous constatons que le test de la médiane est nettement moins puissant (voir le même exemple traité dans la section 3.1.2). Son aptitude à détecter un écart par

rapport à l'hypothèse nulle n'est pas très développé car il exploite de manière très parcellaire les données. Seul le positionnement des individus par rapport à la médiane est utilisé.

3.4 Tests pour les alternatives ordonnées (modèle de localisation)

3.4.1 Position du problème

L'hypothèse alternative du test de Kruskal-Wallis indique qu'un des paramètres de localisation est différent des autres. Il ne donne aucune indication en revanche sur le sens des écarts. Les tests pour hypothèses alternatives permet de pallier cet inconvénient en précisant l'ordre, défini a priori, des paramètres de localisation. Les applications pratiques sont nombreuses : vérifier si, en augmentant les doses d'un médicament, on agit plus efficacement sur la maladie ; est-ce qu'en diminuant le prix d'un produit, on augmente sa part de marché dans les points de vente ; etc.

Le test d'hypothèses s'écrit maintenant

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_K$$

$$H_1 : \theta_1 \leq \theta_2 \leq \dots \leq \theta_K$$

Le rejet de l'hypothèse nulle indique qu'au moins une des inégalités de l'hypothèse alternative est stricte. Notons que dans le cadre d'une comparaison pour $K = 2$ populations, ce type de test est facilement mis en place en spécifiant une hypothèse alternative unilatérale.

Attention, il faut que **l'ordre des paramètres de localisation soit défini à l'avance, par les conditions de l'expérimentation**, par les connaissances du domaine. Il ne s'agit surtout pas de réaliser d'abord les calculs et de définir le test en s'appuyant sur l'ordre des rangs moyens par exemple. Le risque associé au test serait faussé, engendrant une conclusion erronée.

Lorsque l'hypothèse nulle est rejetée, nous savons qu'il y a une relation monotone entre les paramètres de localisation θ_k et le niveau de traitement k . En revanche, nous n'avons pas d'informations sur la forme de la relation (linéaire ou non). En effet, les niveaux sont uniquement des codes qui permettent d'ordonner les traitements. Par convention, nous les noterons $k = 1, 2, \dots, K$, sans que ces valeurs aient une signification particulière.

Enfin, il est possible de généraliser le test pour alternatives ordonnées en définissant des sens opposés des inégalités de part et d'autre d'une valeur centrale. L'hypothèse alternative s'écrit $H_1 : \theta_1 \leq \theta_2 \leq \dots \leq \theta_m \geq \dots \geq \theta_{K-1} \geq \theta_K$. On parle de *test du parapluie*¹¹. Une solution simple consiste alors à décomposer le test en deux sous-tests indépendants, l'une pour l'aile gauche, l'autre pour l'aile droite des inégalités.

¹¹ *Umbrella test* est quand même une appellation autrement plus poétique. Voir <http://de.wikipedia.org/wiki/Umbrella-Test>, c'est en allemand certes, mais il suffit de se focaliser sur les formules.

3.4.2 Test de Jonckheere-Terpstra pour échantillons indépendants

Principe, statistique de test et région critique

Le test de Jonckheere-Terpstra¹² repose sur la statistique

$$J = \sum_{k=1}^{K-1} \sum_{l=k+1}^K U_{kl} \quad (3.7)$$

ou U_{kl} est la statistique de Mann et Whitney définie sur le sous échantillon $(\Omega_k \cup \Omega_l)$

$$U_{kl} = \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} I(x_{ik} < x_{il}) \quad (3.8)$$

$I(x_{ik} < x_{il})$ est une fonction indicatrice définie de la manière suivante

$$I(x_{ik} < x_{il}) = \begin{cases} 1 & \text{si } x_{ik} < x_{il} \\ 0 & \text{sinon} \end{cases}$$

U_{kl} est forcément ≥ 0 . Elle sera d'autant plus grande que les valeurs associées au traitement l ont tendance à être supérieures à celles de k . Au maximum, si toutes les valeurs relatives au groupe l sont plus grandes, nous aurons $U_{kl} = n_k \times n_l$.

De fait, si l'ordonnancement des traitements $k = 1, 2, \dots, K$ correspond effectivement à une augmentation du paramètre de localisation θ_k , la statistique J aura tendance à prendre une valeur élevée. On rejette l'hypothèse nulle si J excède une valeur critique lue dans la table de Jonckheere pour un risque α et une configuration (n_1, \dots, n_K) donnée (voir [13], page 357, table P).

Pour $K = 2$, le test de Jonckheere-Terpstra est équivalent au test de Wilcoxon-Mann-Whitney. Par rapport au test de Kruskal-Wallis, il est plus puissant pour les hypothèses alternatives où il y a effectivement un ordonnancement des paramètres de localisation.

Remarque 12 (Traitement des ex-aequo). S'il y a des individus ex-aequo c.-à-d. 2 ou plusieurs individus appartenant à des groupes différents présentent une valeur identique, le plus simple est de redéfinir la fonction indicatrice de la manière suivante

$$I(x_{ik} < x_{il}) = \begin{cases} 1 & \text{si } x_{ik} < x_{il} \\ 0 & \text{si } x_{ik} > x_{il} \\ \frac{1}{2} & \text{si } x_{ik} = x_{il} \end{cases}$$

Cela n'est pas sans conséquence, la variance de la statistique notamment devra être corrigée. Nous verrons cela plus loin.

¹² Voir http://hdelboy.club.fr/Nonparam.htm#14._Test_de_Jonckheere_-_Tersprat

Remarque 13 (Passage par les rangs). La statistique de Mann et Whitney peut être également obtenue par la somme de rangs (voir section 2.1.7). Pour calculer la quantité U_{kl} , une solution alternative serait de travailler sur la réunion des sous-échantillon $(\Omega_k \cup \Omega_l)$, d'effectif $(n_k + n_l)$. Les rangs d'observations r_{ik} et r_{il} y sont comptabilisés, la somme des rangs pour le groupe Ω_l équivaut à

$$S_l = \sum_{i=1}^{n_l} r_{il}$$

On en déduit la statistique de Mann et Whitney

$$U_{kl} = S_l - \frac{n_l(n_l + 1)}{2}$$

Lorsqu'il y a des ex-aequo dans les données, nous utilisons la règle des rangs moyens.

Approximation pour les grands échantillons

Lorsque les effectifs sont suffisamment grands, la distribution de J sous H_0 peut être approximée par la loi normale de paramètres

$$E(J) = \frac{1}{4} \left(n^2 - \sum_{k=1}^K n_k^2 \right) \quad (3.9)$$

$$V(J) = \frac{1}{72} \left[n^2(2n + 3) - \sum_{k=1}^K n_k^2(2n_k + 3) \right] \quad (3.10)$$

Le test est unilatéral, la région critique au risque α est définie comme suit

$$R.C. : Z = \frac{J - E(J)}{\sqrt{V(J)}} \geq u_{1-\alpha}$$

D'autres tests sont possibles :

- Pour un test unilatéral à gauche, les valeurs diminuent à mesure que le niveau du traitement augmente c.-à-d. la relation est négative, la région critique est définie par $Z \leq u_{1-\alpha}$.
- Pour un test bilatéral, il y a une association entre le niveau et la mesure, la région critique s'écrit $|Z| \geq u_{1-\alpha/2}$.

Exemple 27 (Goût salé et mélange sucré-salé). L'objectif de cette étude (pour plus de détails, voir [13], page 218 à 221) est de montrer que la perception du goût salé est fonction de la proportion relative du sel et du sucre. $K = 4$ niveaux ont été définis, la variable d'intérêt est la perception du goût salé évaluée par des goûteurs.

Il y a plusieurs manières d'organiser les calculs dans un tableur : utiliser les rangs, comparer les valeurs 2 à 2. Le plus simple après réflexion est de passer par un tableau de contingence croisant, en ligne les différentes valeurs de la variable d'intérêt, en colonne les niveaux de traitement (Figure 3.12) :

Fig. 3.12. Test de Jonckheere-Terpstra pour échantillons indépendants - Le goût salé

- Les données sont en colonnes **B** (variable d'intérêt) et **C** (indicateur de niveau de traitement). Nous disposons de $n = 35$ observations.
- A l'aide de l'outil "Tableaux croisés dynamiques", nous élaborons le tableau de contingence entre les valeurs de X et les niveaux. Nous remarquons au passage qu'il n'y a pas d'ex-aequo dans nos données.
- Sur la dernière ligne du tableau, nous avons les effectifs conditionnels, $n_1 = 12$, $n_2 = 9$, $n_3 = 8$ et $n_4 = 6$.
- Il faut maintenant calculer les U_{kl} . Concentrons-nous sur le premier cas U_{12} comptabilisée dans la colonne **L** de la feuille EXCEL.
- La première observation du 2nd groupe est $x_{12} = 13.53$. Il y a 2 observations du 1^{er} groupe qui lui sont inférieures ($x_{11} = 8.82$ et $x_{21} = 11.27$).

- La seconde observation du 2nd groupe est $x_{22} = 28.42$, il y a 5 observations du 1^{er} groupe qui lui sont inférieures.
- En procédant ainsi, on obtient finalement

$$U_{12} = 2 + 5 + 7 + 7 + 8 + 8 + 8 + 10 + 11 = 66$$

- Nous faisons de même pour tous les couples (k, l) de groupes à opposer, nous formons la statistique du test

$$J = 66 + 73 + 62 + 52 + 48 + 36 = 337$$

- L'espérance mathématique sous H_0 est

$$E(J) = \frac{1}{4} \left(n^2 - \sum_{k=1}^K n_k^2 \right) = \frac{1}{4} [35^2 - (12^2 + 9^2 + 8^2 + 6^2)] = 225.0$$

- Et la variance

$$\begin{aligned} V(J) &= \frac{1}{72} \left[n^2(2n+3) - \sum_{k=1}^K n_k^2(2n_k+3) \right] \\ &= \frac{1}{72} [35^2(2 \times 35 + 3) - (12^2(2 \times 12 + 3) + 9^2(2 \times 9 + 3) + 8^2(2 \times 8 + 3) + 6^2(2 \times 6 + 3))] \\ &= 1140.0 \end{aligned}$$

- On en déduit la statistique centrée et réduite

$$Z = \frac{J - E(J)}{\sqrt{V(J)}} = \frac{337 - 225.0}{\sqrt{1140.0}} = 3.31715$$

- A comparer avec $u_{0.95} = 1.64485$ au risque 5%. Nous nous situons dans la région critique, le passage d'un niveau à l'autre entraîne effectivement une augmentation du goût salé.
- La probabilité critique du test est $p = 0.00045$

Lorsque nous aboutissons au rejet de l'hypothèse alternative, pour déterminer les groupes pour lesquels les paramètres de localisation sont significativement décalés, nous pouvons appliquer les comparaisons multiples (section 3.2). A la différence que nous sommes dans un cadre unilatéral, nous ne testons que les $(K-1)$ écarts successifs c.-à-d. $H_1 : \theta_k \leq \theta_{k+1}$.

Remarque 14 (Correction de la variance pour les ex-aequo).

Lorsqu'il y a des ex-aequo, outre le fait d'utiliser les rangs moyens ou de redéfinir la fonction indicatrice $I(\cdot)$, il faut corriger la variance. La nouvelle formule est véritablement diabolique, propre à vous dégoûter des statistiques (voir [3], page 229). Peut être la meilleure manière de la calculer est de passer par les tableaux de contingence et d'utiliser la formulation proposée sur le site du logiciel SAS (voir <http://v8doc.sas.com/sashtml/stat/chap28/sect25.htm>). Quoiqu'il en soit, la correction est assez faible, elle n'est vraiment sensible que si la proportion d'ex-aequo est élevée dans l'échantillon et/ou si un grand nombre d'observations partagent la même valeur.

3.4.3 Test de Page pour échantillons indépendants

Le test de Page est une alternative tout à fait crédible du test de Jonckheere-Terpstra. Pourtant, il en est très peu fait mention dans la littérature. Peut être parce qu'on le retrouve rarement dans les logiciels, peut être aussi parce qu'on le confond avec l'autre test de Page pour les échantillons appariés. Toujours est-il que je l'ai trouvé uniquement dans la bible des tests non paramétriques, et encore sous forme d'exercice (voir [3], page 251, exercice II.6).

La statistique du test repose sur les moyennes de rang conditionnelles, judicieusement combinés avec les numéros de niveau. Les calculs sont autrement plus simples. La statistique s'écrit

$$L = \frac{1}{\sqrt{n}} \sum_{k=1}^K \left(k - \frac{K+1}{2}\right) \left(\bar{r}_k - \frac{n+1}{2}\right) \quad (3.11)$$

Si les paramètres de localisation sont effectivement ordonnés selon les niveaux, on s'attend à ce que lorsque k est inférieur à $\frac{K+1}{2}$ (qui correspond au niveau moyen), \bar{r}_k est aussi inférieur à $\frac{n+1}{2}$ (qui correspond à la moyenne globale des rangs). Le produit est positif. De même, lorsque $k > \frac{K+1}{2}$, alors $\bar{r}_k > \frac{n+1}{2}$, le produit est encore positif. Tout cela concourt à une valeur de la statistique L élevée.

S'il y a des inversions, l'ordre des niveaux n'est pas cohérent avec l'ordre des paramètres de localisation, le produit peut être négatif, réduisant d'autant L .

La véritable différence avec le test de Jonckheere-Terpstra est que nous assumons une forme linéaire de la relation entre k et \bar{r}_k dans l'hypothèse alternative du test. La statistique utilisée (équation 3.11) traduit clairement cette idée. On reconnaît une sorte de covariance entre le numéro de niveau et le rang moyen par groupe.

Sous H_0 , on montre que

$$E(L) = 0$$

$$V(L) = \frac{n+1}{12} \sum_{k=1}^K \frac{1}{n_k} \left(k - \frac{K+1}{2}\right)^2$$

Nous pouvons former la statistique centrée et réduite

$$Z = \frac{L}{\sqrt{V(L)}}$$

La région critique au risque α est associée aux grandes valeurs de L , le test est unilatéral

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

Exemple 28 (Le goût salé (bis)). Reprenons notre exemple du mélange sucré-salé (figure 3.12). Traitons le avec le test de Page. L'organisation de la feuille de calcul est autrement plus classique, s'agissant d'un test de rangs (Figure 3.13) :

Données

x	niveau	rang
8.82	1	1
11.27	1	2
15.78	1	4
17.39	1	5
24.99	1	7
39.05	1	9
47.54	1	10
48.85	1	13
71.66	1	18
72.77	1	19
90.38	1	31
103.13	1	34
13.53	2	3
28.42	2	8
48.11	2	11
48.64	2	12
51.4	2	14
59.91	2	15
67.98	2	17
79.13	2	24
103.05	2	33
19.23	3	6
67.83	3	16
73.68	3	21
75.22	3	22
77.71	3	23
83.67	3	25
86.83	3	28
93.25	3	32
73.51	4	20
85.25	4	26
85.82	4	27
88.88	4	29
90.33	4	30
118.11	4	35

n	35
K	4

niveau	1	2	3	4
n_k	12	9	8	6
S_k	153	137	173	167
r_barre_k	12.750	15.222	21.625	27.833

a_k	7.875	1.389	1.813	14.750
b_k	0.188	0.028	0.031	0.375

L	4.3655
---	--------

E(L)	0
V(L)	1.8646

Z	3.19697
---	---------

u 0.95	1.64485
--------	---------

p-value	0.00069
---------	---------

Fig. 3.13. Test de Page pour échantillons indépendants - Le goût salé

- Nous disposons de $n = 35$ observations, répartis en $K = 4$ groupes numérotés de 1 à 4. Nous attribuons les rangs aux individus, calculés sur l'ensemble de l'échantillon.
- A l'aide de l'outil "Tableaux croisés dynamiques" d'Excel, nous obtenons, pour chaque groupe, les effectifs n_k , les sommes de rangs S_k et les rangs moyens \bar{r}_k
- Pour obtenir la statistique de test, nous formons une quantité intermédiaire a_k telle que

$$a_k = (k - \frac{K+1}{2})(\bar{r}_k - \frac{n+1}{2})$$

Ainsi, pour le premier groupe, nous avons

$$a_1 = (1 - \frac{4+1}{2})(12.750 - \frac{35+1}{2})$$

- La statistique du test est alors égale à

$$L = \frac{1}{\sqrt{n}} \sum_{k=1}^K a_k = \frac{1}{\sqrt{35}} (7.875 + 1.389 + 1.813 + 14.750) = 4.3655$$

- L'espérance de L sous H_0 est nulle. Pour produire la variance, nous formons la quantité intermédiaire b_k , avec

$$b_k = \frac{1}{n_k} \left(k - \frac{K+1}{2} \right)^2$$

Ainsi, pour le 1^{er} groupe

$$b_1 = \frac{1}{12} \left(1 - \frac{4+1}{2} \right)^2 = 0.188$$

La variance est alors égale à

$$V(L) = \frac{n+1}{12} \sum_{k=1}^K b_k = \frac{35+1}{12} (0.188 + 0.028 + 0.031 + 0.375) = 1.8646$$

– Nous obtenons enfin la statistique centrée et réduite

$$Z = \frac{L}{\sqrt{V(L)}} = \frac{4.3655}{\sqrt{1.8646}} = 3.19697$$

- Au risque 5%, nous la comparons à $u_{0.95} = 1.64485$, nous rejetons l'hypothèse nulle. L'ordonnement des groupes selon k est crédible sur nos données.
- La probabilité critique du test est $p = 0.00069$
- On ne manquera pas de faire le rapprochement avec le test de Jonckheere-Terpstra, où la statistique centrée et réduite Z était égale à 3.31715. Finalement, du moins sur nos données, les deux tests produisent des résultats très similaires, avec des conclusions cohérentes. A l'avantage du test de Page, une procédure de calcul autrement plus simple à mettre en place. A son désavantage, l'hypothèse non dite de linéarité d'évolution des paramètres de localisation en fonction de k . Elle n'est pas réhibitoire, loin s'en faut, mais elle constitue une contrainte dont il faut avoir conscience si d'aventure les deux procédures donnent des résultats par trop différents.

Tests de rang dans un modèle d'échelle

Les tests présentés dans les chapitres dédiés aux modèles de localisation sont subordonnés à l'égalité des variabilités (pour simplifier on peut parler de variance ou d'échelle) 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 paramètres de localisation peuvent se révéler robustes. De plus, des tests spécifiques robustes existent également (section 2.4).

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.

Pour la comparaison de $k = 2$ populations, le test d'hypothèses s'écrit

$$\begin{aligned}H_0 : F_1(x) &= F_2(x/\tau), \quad \tau = 1 \\H_1 : F_1(x) &= F_2(x/\tau), \quad \tau \neq 1\end{aligned}$$

τ est le paramètre d'échelle. Il indique la variabilité des valeurs dans les sous populations. Bien évidemment, nous pouvons définir des tests unilatéraux $H_1 : \tau < 1$ (resp. $H_1 : \tau > 1$) c.-à-d. on suppose que la variabilité de la première distribution est inférieure (resp. supérieure) à la seconde.

Si la variance existe, le test bilatéral peut s'écrire

$$\begin{aligned}H_0 : \sigma_1^2 &= \sigma_2^2 \\H_1 : \sigma_1^2 &\neq \sigma_2^2\end{aligned}$$

Les tests présentés dans ce chapitre conviennent pour des variables quantitatives ou ordinales, pourvu que l'on puisse calculer des rangs. On fait l'hypothèse que la distribution sous jacente de la variable d'intérêt est continue.

Une contrainte forte s'applique à la majorité des méthodes présentées dans ce chapitre, les paramètres de localisation doivent être identiques dans les sous-populations. Si ce n'est

pas le cas, le rejet de l'hypothèse nulle peut être imputable à d'autres raisons que les différences de dispersion. Il faudra trouver des solutions appropriées pour y remédier, nous en rediscuterons lors de la présentation effective des tests.

Parmi les tests paramétriques que l'on pourrait placer en face des méthodes présentées dans ce chapitre figurent les très populaires tests de Fisher ($K = 2$) et de Bartlett ($K \geq 2$). Leur défaut est de reposer un peu trop fortement sur la normalité des données. D'autres tests paramétriques plus robustes existent, notamment "les" tests de Levene¹.

4.1 Test de Mood

4.1.1 Principe, statistique de test

Le test de Mood repose classiquement sur une statistique de rang linéaire. Sa principale astuce est de proposer une fonction score telle que sa valeur est faible près de la moyenne des rangs $\frac{n+1}{2}$, élevée lorsqu'on s'en écarte. Les valeurs extrêmes se voient ainsi attribuées des scores élevés, ce qui transparaît dans la statistique de test.

Sans restreindre la généralité du propos, nous considérerons que le 1^{er} groupe sert de référence ($n_1 < n_2$; si $n_1 = n_2$, la somme des scores de Ω_1 est inférieure à celle de Ω_2).

La statistique du test de Mood s'écrit

$$M = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\frac{r_{i1}}{n+1} - \frac{1}{2} \right)^2 \quad (4.1)$$

On retrouve dans certaines publications une autre formulation, complètement équivalente

$$M' = \sum_{i=1}^{n_1} \left(r_{i1} - \frac{n+1}{2} \right)^2 \quad (4.2)$$

Le score $s_i = \left(\frac{r_{i1}}{n+1} - \frac{1}{2} \right)^2$ répond exactement au comportement désiré. Il est minimal lorsque l'on se situe vers les valeurs centrales (la moyenne des rangs). Il est maximal aux extrémités de l'échantillon (rang proche de 1 ou rang proche de n).

La région critique est du domaine des grandes valeurs de M . Lorsqu'elle est supérieure au seuil critique lue dans la table, on rejette l'hypothèse nulle.

¹ Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf

4.1.2 Approximation par la loi normale

Réaliser ce test sur des petits effectifs est tributaire de la disponibilité des tables statistiques qui sont, malheureusement, quasiment introuvables. Heureusement, lorsque les effectifs augmentent, il est très rapidement ($n \geq 30$) possible de passer à l'approximation normale. Les paramètres de la distribution de M sous H_0 sont

$$E(M) = \frac{n-1}{12(n+1)}$$

$$V(M) = \frac{n_2(n^2-4)}{180n_1(n+1)^3}$$

La statistique centrée réduite Z est définie par

$$Z = \frac{M - E(M)}{\sqrt{V(M)}}$$

La région critique pour un test bilatéral au risque α s'écrit

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

Exemple 29 (Ergonomie des tableaux de bord de véhicules). Un constructeur automobile désire évaluer l'accès à certaines fonctionnalités de son modèle phare. Il a le choix entre proposer une organisation classique "A" des commandes, bien établie dans le segment, ou proposer une organisation intuitive "B", novatrice. Il n'a pas trop d'inquiétude quant au temps d'accès moyen, les 2 dispositifs se valent. Il a des doutes en revanche sur l'appréhension du dispositif intuitif, il pense qu'il aura à gérer deux types de réactions : soit le conducteur comprend tout de suite et arrive à ses fins ; soit il est dérouté et se perd dans les méandres des boutons qui parsèment le tableau de bord. Bref, il souhaite comparer la variance du temps d'accès aux fonctionnalités.

Les données sont disponibles dans une feuille Excel (Figure 4.1) :

- Le fichier comporte $n = 30$ observations, avec $n_1 = 15$ (groupe "A") et $n_2 = 15$ (groupe "B").
- On déduit des données la colonne des rangs, puis les scores $s_i = (\frac{r_{i1}}{n+1} - \frac{1}{2})^2$. Ainsi, pour la première observation appartenant au premier groupe, nous obtenons $s_{11} = (\frac{24}{30+1} - \frac{1}{2})^2 = 0.075$; pour le second $s_{21} = (\frac{8}{30+1} - \frac{1}{2})^2 = 0.059$; etc.
- A partir de là, nous pouvons former la somme des scores conditionnelle (la somme des scores restreinte au 1^{er} groupe)

$$S_1 = \sum_{i=1}^{n_1} s_{i1} = 0.075 + 0.059 + \dots + 0.000 + 0.059 = 0.7802$$

- Et la statistique

$$M = \frac{1}{n_1} S_1 = \frac{1}{15} 0.7802 = 0.0520$$

i	duree	organisation	rang	score	
1	4.953	A	24	0.075	n 30
2	2.524	A	8	0.059	
3	2.207	A	6	0.094	n1 15
4	3.153	A	9	0.044	n2 15
5	4.637	A	21	0.031	
6	4.110	A	18	0.007	S1 0.7802
7	3.607	A	12	0.013	
8	3.910	A	17	0.002	M 0.0520
9	3.521	A	11	0.021	
10	2.404	A	7	0.075	E(M) 0.0780
11	2.112	A	4	0.138	V(M) 0.0002
12	6.947	A	28	0.163	
13	3.857	A	16	0.000	Z -2.0072
14	3.756	A	15	0.000	
15	4.836	A	23	0.059	u _(0.975) 1.9600
16	1.240	B	2	0.190	
17	8.163	B	29	0.190	p-value 0.0447
18	3.755	B	14	0.002	
19	10.460	B	30	0.219	
20	2.172	B	5	0.115	
21	4.672	B	22	0.044	
22	4.376	B	20	0.021	
23	1.002	B	1	0.219	
24	1.855	B	3	0.163	
25	4.227	B	19	0.013	
26	3.727	B	13	0.007	
27	3.409	B	10	0.031	
28	5.973	B	26	0.115	
29	5.786	B	25	0.094	
30	6.331	B	27	0.138	

Fig. 4.1. Test de Mood - Données "Ergonomie"

– Nous calculons l'espérance sous H_0

$$E(M) = \frac{n-1}{12(n+1)} = \frac{30-1}{12(30+1)} = 0.0780$$

Et la variance

$$V(M) = \frac{n_2(n^2-4)}{180n_1(n+1)^3} = \frac{15(30^2-4)}{180 \times 15(30+1)^3} = 0.0002$$

– La statistique centrée et réduite est obtenue avec

$$Z = \frac{0.0520 - 0.0780}{\sqrt{0.0002}} = -2.0072$$

– Pour un test bilatéral à 5%, nous comparons $|Z| = 2.0072$ à $u_{1-0.05/2} = u_{0.975} = 1.9600$. Nous sommes dans la région critique, les dispersions ne sont pas les mêmes dans les sous populations.

– La probabilité critique du test est $p = 0.0447$

– Dans un test unilatéral, $H_1 : \tau < 1$, nous comparerons $Z = -2.0072$ au seuil $u_{0.05} = -1.6449$. On conclura que la disparité des temps d'adaptation face aux boutons du tableau de bord est moins grande lorsque leur disposition est classique.

4.1.3 Correction pour les ex-aequo

Lorsqu'il y a des ex-aequo, nous utilisons le principe des scores moyens c.-à-d. nous effectuons la péréquation des scores pour les individus qui présentent des valeurs identiques. La variance doit être

modifiée pour en tenir compte. Nous ne présentons cependant pas la formule corrigée car on se rend compte plus loin (chapitre 5, page 105) qu'il est possible d'obtenir facilement la statistique centrée et réduite ajustée \tilde{Z} en adoptant une démarche générique.

4.1.4 Lorsque les paramètres de localisation sont différents

Le problème est autrement plus délicat lorsque les paramètres de localisation sont différents d'un groupe à l'autre :

1. S'ils ont connus (les médianes sont connues par exemple), nous pouvons centrer les données dans chacun des groupes avant de constituer l'échantillon global présenté au test. Nous pouvons alors utiliser la démarche habituelle. Les distributions exactes et asymptotiques des statistiques de test sont conservés.
2. S'ils sont inconnus, certains auteurs préconisent d'estimer les paramètres de localisation conditionnels (la médiane empirique par exemple, mais d'autres estimateurs *robustes* peuvent être utilisés), de centrer les données dans chaque groupe avant de présenter l'ensemble au test (voir [1], page 321). D'autres s'y opposent farouchement (voir [13], page 160). Il semble cependant que cette stratégie soit viable pour les grands effectifs. En effet, les distributions asymptotiques des statistiques de test sont conservés (voir [3], page 179). Pour les petits effectifs, il n'y a pas, semble-t-il, d'études réellement convaincantes à ce sujet.

Attention, les logiciels n'entrent pas dans ce genre de considérations et réalisent les calculs à partir des données brutes. S'il y a un centrage à faire autour des médianes conditionnelles dans les groupes, nous devons procéder en amont, lors de la phase de préparation des données.

4.2 Test de Klotz

4.2.1 Principe du test de Klotz

Le test de Klotz repose également sur une statistique de rang linéaire. Il utilise une fonction score qui le rend particulièrement performant lorsque les données se rapprochent de la distribution normale, aussi performant que son homologue paramétrique, le test de Fisher.

Il repose sur la statistique

$$L = \sum_{i=1}^{n_1} \left[\Phi^{-1} \left(\frac{r_{i1}}{n+1} \right) \right]^2 \quad (4.3)$$

où $\Phi^{-1}(\cdot)$ est la fonction inverse de la fonction de répartition de la loi normale.

Les scores sont donc le carré des quantiles de la loi normale. Ils correspondent au carré de ceux de Van der Waerden, comme ceux de Mood correspondaient au carré de ceux de Wilcoxon-Mann-Whitney. Sous cet angle, on perçoit mieux pourquoi le test de Klotz, tout comme le test de Mood, est adapté à

la différenciation des échelles entre les populations; et pourquoi, par rapport au test de Mood, il est particulièrement performant lorsque la distribution sous-jacente des données est gaussienne.

Les valeurs critiques sont tabulées² pour $n_1 + n_2 \leq 20$. Au delà, nous pouvons nous tourner vers la distribution asymptotique sous H_0 . Elle est gaussienne de paramètres

$$E(L) = \frac{n_1}{n} \sum_{i=1}^n \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^2 \quad (4.4)$$

$$V(L) = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^n \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^4 - \frac{n_2}{n_1(n-1)} [E(L)]^2 \quad (4.5)$$

La région critique pour la statistique centrée et réduite Z est du même type que celui de Mood.

Exemple 30 (Retour sur les données "Ergonomie"). Revenons sur les données "Ergonomie des tableaux de bord de véhicules" analysée précédemment (Test de Mood, page 93). La structure de la feuille de calcul est quasi identique (Figure 4.2) :

i	duree	rganisatio	rang	score	score ²	
1	4.953	A	24	0.567	0.321	n 30
2	2.524	A	8	0.422	0.178	n1 15
3	2.207	A	6	0.748	0.560	n2 15
4	3.153	A	9	0.305	0.093	S1 6.4916
5	4.637	A	21	0.212	0.045	L 6.4916
6	4.110	A	18	0.041	0.002	E(L) 12.2016
7	3.607	A	12	0.082	0.007	V(L) 7.1053
8	3.910	A	17	0.015	0.000	Z -2.1421
9	3.521	A	11	0.139	0.019	u (0.975) 1.9600
10	2.404	A	7	0.567	0.321	p-value 0.0322
11	2.112	A	4	1.279	1.636	
12	6.947	A	28	1.690	2.857	
13	3.857	A	16	0.002	0.000	
14	3.756	A	15	0.002	0.000	
15	4.836	A	23	0.422	0.178	
16	1.240	B	2	2.304	5.309	
17	8.163	B	29	2.304	5.309	
18	3.755	B	14	0.015	0.000	
19	10.460	B	30	3.417	11.678	
20	2.172	B	5	0.978	0.957	
21	4.672	B	22	0.305	0.093	
22	4.376	B	20	0.139	0.019	
23	1.002	B	1	3.417	11.678	
24	1.855	B	3	1.690	2.857	
25	4.227	B	19	0.082	0.007	
26	3.727	B	13	0.041	0.002	
27	3.409	B	10	0.212	0.045	
28	5.973	B	26	0.978	0.957	
29	5.786	B	25	0.748	0.560	
30	6.331	B	27	1.279	1.636	

Fig. 4.2. Test de Klotz - Données "Ergonomie"

- A partir des données, nous calculons les rangs puis les scores. Par exemple, pour le premier individu du premier groupe, nous avons $s_{11} = [\Phi^{-1}(\frac{24}{30+1})]^2 = [0.753]^2 = 0.567$; pour le second $s_{21} = [\Phi^{-1}(\frac{8}{30+1})]^2 = [-0.649]^2 = 0.422$; etc.

² Les tables sont rares également pour le test de Klotz, de surcroît configurées pour des situations particulières. Sur ce site par exemple <http://www.win.tue.nl/~markvdw/tabellen.htm>, la table des valeurs critiques du test de Klotz sont disponibles uniquement pour les échantillons équilibrés c.-à-d. $n_1 = n_2 = \frac{n}{2}$.

- Nous calculons la somme des scores restreinte au premier groupe, qui est aussi la statistique du test de Klotz

$$L = 0.567 + 0.422 + \dots + 0.002 + 0.422 = 6.4916$$

- Nous calculons l'espérance sous H_0

$$E(L) = \frac{n_1}{n} \sum_{i=1}^n \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^2 = \frac{15}{30} \times (0.567 + 0.422 + \dots + 0.748 + 1.279) = 12.2016$$

- Pour obtenir la variance, nous avons besoin de la série des $s_i^2 = \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^4$, nous rajoutons la colonne correspondante dans notre tableau, puis nous effectuons la somme $\sum_{i=1}^n s_i^2 = \sum_{i=1}^n \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^4 = 47.324$. Nous pouvons passer au calcul de la variance

$$\begin{aligned} V(L) &= \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^n \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^4 - \frac{n_2}{n_1(n-1)} [E(L)]^2 \\ &= \frac{15 \times 15}{30(30-1)} \times 47.324 - \frac{15}{15(30-1)} [12.2016]^2 \\ &= 7.1053 \end{aligned}$$

- La statistique centrée et réduite s'écrit

$$Z = \frac{L - E(L)}{\sqrt{V(L)}} = \frac{6.4916 - 12.2016}{\sqrt{7.1053}} = -2.1421$$

- Pour un test bilatéral à 5%, le seuil critique est $u_{0.975} = 1.9600$. Comme $|Z| = 2.1421 > 1.9600 = u_{0.975}$, nous pouvons rejeter l'hypothèse nulle d'égalité des dispersions.
- La probabilité critique du test est $p = 0.0322$.

4.2.2 Quelques considérations sur les performances

Les tests de Mood et Klotz se comparent favorablement à leurs homologues paramétriques. Le test de Mood est un peu moins efficace que le test de Fisher lorsque les distributions sous-jacentes des données sont normales. Dans tous les autres cas, il est meilleur. Le test de Klotz fait mieux encore en étant aussi efficace que le test de Fisher, même lorsque les données sont normales.

Le test de Klotz est en général meilleur que le test de Mood (voir [1], page 321), sauf lorsque nous avons affaire à des distributions à queues lourdes. Les comparaisons sur données simulées établissant ces résultats sont détaillées dans [3], pages 174 et 175.

4.2.3 Sorties des logiciels

Le test de Klotz fait partie des tests fondés sur les statistiques de rang linéaires. Les logiciels s'appuient sur ce canevas. De fait, leurs productions sont totalement standardisées. Le passage d'un test à l'autre se fait simplement sur la base d'une modification de la fonction score. Nous montrons dans cette section

la lecture des résultats du test de Klotz, mais les commentaires émis sont valables pour tous les tests similaires (Mood, Ansari-Bradley, etc.). Nous utiliserons les données "Ergonomie" (Figure 4.2).

TANAGRA produit, pour chaque groupe, le nombre d'observations n_k , la somme des scores S_k et la moyenne des scores \bar{r}_k . La statistique de test est élaborée à partir de ces informations (Figure 4.3) :

Klotz Scale Test 1									
Parameters									
Parameters									
Sort results: no									
Results									
Attribute_Y	Attribute_X	Description					Statistical test		
duree	organisation	Value	Examples	Average	Scores sum	Scores mean	Two-Sample Test		
		A	15	3.7689	6.4916	0.4328	S	6.49164	
		B	15	4.4765	17.9117	1.1941	E(S)	12.20165	
							V(S)	7.10530	
		All	30	4.1227	24.4	0.8134	Z	2.14213	
							p-value	0.03218	
								One-way Analysis	
							Chi-Square	4.58873	
							d.f.	1	
							p-value	0.03218	

Fig. 4.3. Test de Klotz - Données "Ergonomie" - TANAGRA

- $S = 6.49164$ correspond à la somme des scores du groupe de référence "Organisation = A".
- Sous H_0 , nous obtenons l'espérance $E(S) = 6.49164$ et la variance $V(S) = 7.10530$ (voir équations, page 96).
- La valeur absolue de la statistique centrée et réduite est formée

$$|Z| = \frac{|6.49164 - 12.20165|}{\sqrt{7.10530}} = 2.14213$$

- La probabilité critique du test est $p = 0.03218$

La statistique basée sur un schéma d'analyse de variance est aussi fournie. Elle doit être cohérente avec la précédente, avec $\chi^2 = Z^2$. Ici, $\chi^2 = 4.58873$, avec la même probabilité critique. Lorsque la variable indépendante définissant les groupes possède un nombre de modalités strictement supérieur à 2, seule cette statistique est fournie.

La production de TANAGRA a été calée sur ceux de **SAS**. Il est donc normal que les sorties concordent en tous points (Figure 4.4).

Enfin, précisons encore une fois que les logiciels ne se préoccupent pas de savoir si les paramètres de localisation sont identiques dans les sous-populations. Si ce n'est pas le cas, nous devons faire un travail préalable de centrage des données, à l'intérieur de chaque groupe, avant de lancer le test.

The NPAR1WAY Procedure					
Klotz Scores pour la variable durees					
Classée par variable organisation					
organisation	Nb	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
A	15	6.491636	12.201650	2.665576	0.432775
B	15	17.911664	12.201650	2.665576	1.194111
Klotz Two-Sample Test					
Statistic			6.4916		
Z			-2.1421		
One-Sided Pr < Z			0.0161		
Two-Sided Pr > Z			0.0322		
Klotz One-Way Analysis					
Chi-Square			1.5887		
DF			1		
Pr > Chi-Square			0.0322		

Fig. 4.4. Test de Klotz - Données "Ergonomie" - SAS

4.3 Test robuste de Moses

Ce test se distingue principalement par le fait qu'il ne requiert pas l'égalité des paramètres de localisation conditionnels. Une restriction forte est levée. Il ne repose pas à proprement parler sur une statistique de rang linéaire. Nous dirons plutôt qu'il utilise, dans sa seconde partie tout du moins, un test de basé sur les rangs, le test de Wilcoxon-Mann-Whitney en l'occurrence, pour établir le rejet ou non de l'hypothèse nulle. Préalablement, il est nécessaire de réaliser des calculs sur les données brutes pour produire les valeurs intermédiaires que l'on présentera au test de Wilcoxon. Ce test ne fonctionne que sur les données quantitatives, lorsque l'écart entre 2 valeurs est interprétable.

En anglais, ce test est désigné par l'appellation "Moses rank-like test for scale differences" (voir [13], pages 161 à 166 ; notre présentation s'en inspire directement). Nous ne devons pas le confondre avec le test de réaction extrêmes de Moses ("Moses test of extreme reaction")³, destiné à déterminer le surcroît d'apparition de valeurs dans les queues de distribution d'un groupe expérimental (des deux côtés), par rapport à un groupe de contrôle.

4.3.1 Construction du test

Première étape : construction des données intermédiaires

Dans un premier temps, à partir des données originelles, nous devons construire des données intermédiaires qui traduisent la variabilité dans les sous-populations. On procède de la manière suivante.

Les données sont subdivisées aléatoirement en blocs⁴ de taille h . Les individus inclus dans un bloc doivent appartenir à la même sous-population.

³ voir <http://www2.chass.ncsu.edu/garson/PA765/mann.htm>

⁴ Dénomination un peu alambiquée, j'en conviens, à ne pas confondre avec les blocs pour les échantillons appariés. Utilisons ce mot afin qu'il n'y ait pas d'ambiguïtés entre ces nouveaux sous-ensembles et les groupes relatifs aux sous-populations à comparer.

Prenons un exemple simple, nous disposons d'un échantillon de taille $n = 46$ observations, dont $n_1 = 25$ appartiennent à la première sous-population et $n_2 = 21$ à la seconde. On souhaite définir des groupes de taille $h = 5$. Nous créons $m_1 = n_1 \div h = 25 \div 5 = 5$ blocs de la 1^{ère} sous population, et $m_2 = n_2 \div h = 21 \div 5 = 4$ de la 2^{nde}. Lorsque la division n'est pas entière, les individus en surplus sont simplement ignorés dans le reste de l'étude.

Pour chaque bloc, nous produisons un indicateur de sa variabilité interne en calculant la somme des carrés des écarts à la moyenne (du bloc). La formule pour le bloc n^{ol} s'écrit

$$v_l = \sum_{j=1}^h (x_{jl} - \bar{x}_l)^2$$

où \bar{x}_l est la moyenne de la variable d'intérêt à l'intérieur du bloc.

Nous disposons ainsi de m nouvelles observations synthétiques, dont m_1 sont relatives à la première sous population, m_2 à la seconde. Si la dispersion est plus forte (resp. moins forte) dans la première sous population, on s'attend à ce qu'en moyenne, les valeurs v_l associées soient plus élevées (resp. plus faibles). C'est ce qu'essaie de mettre en évidence le test qui suit.

Deuxième étape : test de Wilcoxon-Mann-Whitney sur les données intermédiaires

Partant de l'idée que les valeurs intermédiaires v_l sont des indicateurs de variabilité, et qu'elles sont associées chacune à une des sous population. L'étape suivante consiste tout simplement à comparer les caractéristiques de tendance centrale de la distribution de v_l dans les sous populations. Nous pourrions utiliser un test de comparaison de moyennes d'ailleurs, mais en l'absence d'informations sur la distribution sous-jacente des v_l , mieux vaut nous tourner vers un test non paramétrique, qui plus est adapté aux petits échantillons car il est évident que nous ne disposerons pas de beaucoup d'observations v_l (m sera assez petit en général, en réalité il y a un compromis à trouver, voir les commentaires ci-dessous).

Le schéma du test de Wilcoxon-Mann-Whitney n'est pas modifié. A la différence que nous disposons de $m = m_1 + m_2$ observations. Ce sont les valeurs v_l qui sont transformés en rangs r_l . La définition de la statistique de test et de la région critique est strictement la même (voir 2.1, page 34).

Exemple 31 (Retour sur l'ergonomie des tableaux de bord). Revenons sur les données "Ergonomie des tableaux de bord de véhicules" analysée précédemment (Test de Mood, page 93 ; Test de Klotz, page 96). Le travail préparatoire est un peu plus important, la structure de la feuille de calcul s'en ressent (Figure 4.5) :

- Nous disposons de $n = 30$ observations, répartis en 2 groupes avec $n_1 = 15$ et $n_2 = 15$. Nous souhaitons comparer la dispersions dans les groupes.
- Tout d'abord, nous mélangeons aléatoirement les données à l'intérieur de chaque groupe Ω_k . Nous pouvons constater cela en observant les numéros d'individus dans la colonne i qui ne sont plus ordonnés maintenant.
- Nous choisissons $h = 3$, nous produirons donc $m_1 = 15 \div 3 = 5$ blocs du 1^{er} groupe, et $m_2 = 15 \div 3 = 5$ du second.

i	duree	organisation	h	3
11	2.112	A		
8	3.910	A		
4	3.153	A		
15	4.836	A		
6	4.110	A		
1	4.953	A		
13	3.857	A		
10	2.404	A		
12	6.947	A		
14	3.756	A		
3	2.207	A		
5	4.637	A		
2	2.524	A		
7	3.607	A		
9	3.521	A		
30	6.331	B		
17	8.163	B		
24	1.855	B		
16	1.240	B		
19	10.460	B		
29	5.786	B		
23	1.002	B		
26	3.727	B		
27	3.409	B		
28	5.973	B		
21	4.672	B		
18	3.755	B		
22	4.376	B		
25	4.227	B		
20	2.172	B		

Blocs	l	v _l	organisation	Rang
2.112 3.910 3.153	1	1.630	A	3
4.836 4.110 4.953	2	0.417	A	1
3.857 2.404 6.947	3	10.766	A	8
3.756 2.207 4.637	4	3.027	A	5
2.524 3.607 3.521	5	0.725	A	2
6.331 8.163 1.855	6	21.061	B	9
1.240 10.460 5.786	7	42.507	B	10
1.002 3.727 3.409	8	4.440	B	7
5.973 4.672 3.755	9	2.484	B	4
4.376 4.227 2.172	10	3.034	B	6

m	10
m1	5
m2	5
W _s	19
E(W _s)	27.500
V(W _s)	22.917
Z	-1.776
u (0.975)	1.960
p-value	0.0758

Fig. 4.5. Test de Moses - Données "Ergonomie"

- A partir des blocs, nous calculons les indicateurs v_l . Voyons en le détail sur la première ligne $l = 1$, relatif au premier groupe Ω_1 ("Organisation = A"). Nous disposons des valeurs $\{2.112, 3.910, 3.153\}$. La moyenne afférente est $\bar{x}_{l=1} = \frac{2.112+3.910+3.153}{3} = 3.058$. Nous pouvons former v_1 , avec

$$v_1 = (2.112 - 3.058)^2 + (3.910 - 3.058)^2 + (3.153 - 3.058)^2 = 1.630$$

- De fait, nous disposons d'un nouveau tableau de données intermédiaires avec l , v_l et leurs groupes d'appartenance respectifs (la colonne "Organisation"). Le test de Wilcoxon est mis en oeuvre à partir de ce nouveau tableau.
- Nous calculons la colonne des rangs. Nous produisons alors la statistique de Wilcoxon (pour varier les plaisirs, nous appliquons le test original de Wilcoxon décrit en section 2.1.6, page 44).
- La statistique est obtenue par la somme des rangs du 1^{er} groupe, soit :

$$W_s = 3 + 1 + 8 + 5 + 2 = 19$$

- Pour déterminer si nous situons dans la région critique pour un test bilatéral à 5%, nous pouvons consulter la table de Wilcoxon (voir <http://courses.wcupa.edu/rbove/eco252/252suppkey.htm>). Elle nous indique que nous devons rejeter l'hypothèse nulle si ($W_s \leq 17$ ou $W_s \geq 38$). Ce qui n'est pas le cas pour nous, les données sont compatibles avec l'hypothèse d'égalité des dispersions dans les sous-populations.
- A titre indicatif, nous avons utilisé l'approximation normale. Les paramètres de la distribution sous H_0 sont

$$E(W_s) = \frac{m_1(m_1 + m_2 + 1)}{2} = \frac{5(5 + 5 + 1)}{2} = 27.500$$

Et

$$V(W_s) = \frac{1}{12}(m_1 + m_2 + 1)m_1m_2 = \frac{1}{12}(5 + 5 + 1) \times 5 \times 5 = 22.917$$

– La statistique centrée réduite est

$$Z = \frac{(19 - 27.500)}{\sqrt{22.917}} = -1.776$$

Puisque $|Z| = 1.776 < u_{1-\alpha/2} = u_{0.975} = 1.960$, nous aboutissons à la même conclusion, nous ne pouvons pas rejeter l'hypothèse nulle.

– Toujours avec l'approximation normale, la probabilité critique du test est $p = 0.0758$.

De manière générale, le test de Moses est moins puissant que les test de Klotz ou de Mood. En revanche, il a l'avantage de s'appliquer même lorsque les paramètres de localisation conditionnels sont différents. En ce sens, il nous affranchit d'un travail de vérification préalable lors de l'appréhension des données.

4.3.2 Quelques commentaires

Premier commentaire important, $h \geq 2$ forcément. En y regardant de plus près on se rend compte que le choix de h influe sur la qualité des résultats : si h est grand, nous disposerons d'indicateurs de variabilité dans les blocs v_l de bonne qualité, mais le nombre de valeurs m que nous présenterons au test de Wilcoxon-Mann-Whitney sera faible, mettant en péril la teneur des conclusions de celui-ci. A l'inverse, si nous diminuons h , m sera grand, le test subséquent produira des résultats fiables, mais sur des valeurs v_l qui sont, elles, sujettes à caution car estimées sur trop peu d'observations.

Second commentaire, la démarche est facilement généralisable à la comparaison de la dispersion dans $K \geq 2$ populations. La constitution des données intermédiaires v_l ne pose aucun problème, et il suffit de substituer le test de Kruskal-Wallis au test de Wilcoxon-Mann-Whitney.

Enfin, troisième commentaire, ce test est caractéristique de l'idée sous-jacente derrière les tests de comparaison de dispersion. Au final, nous devons aboutir à une comparaison de caractéristique de localisation. Pour que cette dernière revienne à une comparaison des dispersions dans la représentation initiale des données x_i , nous devons les transformer de manière judicieuse. Les test de Mood et Klotz utilisent des fonctions scores s_i calculés à partir des rangs, le test de Moses s'appuie sur un procédé plus original pour produire les quantités v_l .

4.4 Généralisation à $K \geq 2$ groupes

La généralisation des tests basés sur les statistiques de rang linéaires à la comparaison de $K \geq 2$ populations ne pose aucune problème. On se place dans ce cas dans le canevas d'une analyse de variance sur les scores, charge à ces dernières de mettre en évidence des disparités de paramètres de localisation ou des paramètres d'échelle.

Les tests de Mood et de Klotz peuvent être généralisés. Nous ne détaillons pas les formules dans la mesure où nous présenterons l'approche générique plus loin (chapitre 5, page 105). Ces tests en tous les cas constituent des alternatives crédibles aux tests paramétriques de comparaison de variance tels que le test de Bartlett ou le test de Levene. La contrainte forte reste toujours que les paramètres de localisation doivent être identiques dans les sous populations.

A titre d'exemple, nous montrons les sorties du logiciel SAS pour le test de Mood sur nos données "Ergonomie" (Figure 4.6). La statistique de test est basée sur la formulation M' (équation 4.2), elle est différente de la notre (figure 4.1). Les statistiques Z et probabilités critiques sont en revanche bien identiques (ouf!). SAS affiche en plus, automatiquement, le calcul basé sur un schéma d'analyse de variance sur les scores, qui est applicable pour $K \geq 2$. Elle est distribuée selon la loi du χ^2 lorsque les effectifs sont suffisamment grands. Dans notre cas, $\chi^2 = Z^2$ puisque $K = 2$. Lorsque $K > 2$, seule la seconde partie des résultats est disponible. Ce qui est tout à fait cohérent.

The NPARIWAY Procedure					
Mood Scores pour la variable duree					
Classée par variable organisation					
organisation	Nb	Somme des scores	Attendue sous H0	Écart-type sous H0	Score moyen
A	15	749.750	1123.750	186.333035	49.983333
B	15	1497.750	1123.750	186.333035	99.850000
Mood Two-Sample Test					
Statistic			749.7500		
Z			-2.0072		
One-Sided Pr < Z			0.0224		
Two-Sided Pr > Z			0.0447		
Mood One-Way Analysis					
Chi-Square			4.0287		
DF			1		
Pr > Chi-Square			0.0447		

Fig. 4.6. Sorties SAS, test de Mood - Données "Ergonomie"

Retour sur les statistiques de rang linéaires

5.1 Statistiques de rang linéaires

La grande majorité des tests présentés dans cette partie de notre support ont pour point commun d'être basés sur les statistiques de rangs. Nous écrirons la somme des rangs du groupe Ω_k de la manière générique suivante

$$S_k = \sum_{i=1}^n c_{ik} f(r_i)$$

où c_{ik} est une variable indicatrice qui vaut 1 si l'individu n^oi appartient à l'échantillon Ω_k , 0 sinon.

Nous pouvons simplifier l'écriture en écrivant

$$S_k = \sum_{i=1}^{n_k} f(r_{ik})$$

sachant que les rangs sont toujours calculés par rapport à l'échantillon global Ω .

Tous les tests abordés dans cette partie utilise les quantités S_k , elles sont combinées de différentes manières selon que l'on cherche à comparer $K = 2$ ou $K > 2$ échantillons, mais fondamentalement la démarche est toujours la même.

Dans ce contexte, on se rend compte que la fonction score $f(r_i)$ joue un rôle crucial. En fait, elle détermine le type d'information que l'on cherche à mettre en exergue (décalage des paramètres de localisation ou inégalité des paramètres d'échelle), et elle doit être adaptée aux spécificités de la distribution des données (symétrie, asymétrie, importance des queues de distributions).

Dans ce qui suit nous allons ré-écrire de manière générique tous les tests non paramétriques de comparaison de population basés sur des statistiques de rang. L'intérêt est que, par la suite, en introduisant la bonne fonction score, nous retrouvons tous les tests que nous avons décrit précédemment dans ce support, que ce soit pour le cas $K = 2$ ou $K \geq 2$. Nous n'abordons que les procédures asymptotiques dans ce chapitre.

Remarque 15 (Un délice à programmer). Pour un informaticien, cette situation est le rêve absolu. Dans la hiérarchie de classes, il suffit d'écrire les classes ancêtres de la manière la plus générique possible. En bout de ligne, pour chaque classe de calcul relative à un test, la seule méthode à surcharger est la fonction score. Ainsi, on pourra aligner un grand nombre de techniques dans un logiciel avec un minimum d'effort. C'est ce que nous avons fait d'ailleurs dans le logiciel TANAGRA. Vive la modélisation objet !

5.2 Écriture des statistiques de test pour ($K = 2$)

Les sommes des scores sont calculées selon les formules décrites ci-dessus¹. Pour fixer les idées, nous dirons que Ω_1 est le groupe de référence, avec $n_1 < n_2$. Nous définissons

$$S = S_1$$

Pour produire la statistique Z , qui est distribuée asymptotiquement selon une loi normale centrée et réduite, nous calculons

$$Z = \frac{S - E_0(S)}{\sqrt{V_0(S)}} \quad (5.1)$$

$E_0(S)$ et $V_0(S)$ sont respectivement l'espérance mathématique et la variance de S sous l'hypothèse nulle H_0 :

$$E_0(S) = \frac{n_1}{n} \sum_{i=1}^n s_i \quad (5.2)$$

et

$$V_0(S) = \frac{n_1 n_2}{n(n-1)} \left[\sum_{i=1}^n (s_i - \bar{s})^2 \right] \quad (5.3)$$

où

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i$$

Pour un test bilatéral, on utilisera $|Z|$. Si besoin est, on pourra introduire la correction de continuité.

¹ <http://v8doc.sas.com/sashtml/stat/chap47/sect15.htm>

5.3 Écriture des statistiques de test pour ($K \geq 2$)

Pour ($K \geq 2$)², nous sommes dans un schéma d'analyse de variance³. L'hypothèse nulle est que les caractéristiques de tendance centrale des distributions conditionnelles sont toutes égales. La statistique de test traduit l'idée du rapport entre variances si l'on fait le parallèle avec l'approche paramétrique. Elle s'écrit

$$C = \frac{\sum_{k=1}^K \frac{1}{n_k} \times [S_k - E_0(S_k)]^2}{T^2} \quad (5.4)$$

$E_0(S_k)$ est l'espérance de la somme des scores pour le groupe $n^o k$ sous l'hypothèse nulle

$$E_0(S_k) = n_k \bar{s}$$

où \bar{s} est la moyenne globale des scores

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i$$

T^2 est la variance totale des scores, estimée sur l'échantillon global

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

Sous H_0 , C suit une loi du χ^2 à $(K - 1)$ degrés de liberté. La région critique du test correspond aux valeurs anormalement élevées. Au risque α :

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

5.4 Les principales fonctions scores

Les procédures génériques de test étant décrites, nous pouvons énumérer maintenant les principales fonctions scores⁴ et les tests associés. Pour les situer simplement, il y a principalement deux catégories de fonctions :

1. Les fonctions qui traduisent le positionnement relatif des points, les rangs eux même, ou les transformations monotones des rangs. Ces fonctions servent aux tests de comparaison des caractéristiques de tendance centrale (de localisation). On y retrouve les fonctions score de Wilcoxon, de Van der Waerden, de la médiane.

² Si $K = 2$, les résultats de la procédure proposée dans cette section sont exactement équivalents à ceux de la procédure précédente.

³ <http://v8doc.sas.com/sashtml/stat/chap47/sect16.htm>

⁴ <http://v8doc.sas.com/sashtml/stat/chap47/sect17.htm>

2. Les fonctions qui traduisent l'écartement des observations autour d'une position centrale. Elles ont pour particularité de produire une valeur élevée lorsque l'observation est éloignée de la valeur centrale, une valeur faible lorsqu'elle en est proche. De manière schématique (et imagée), nous dirons que la fonction est de forme parabolique à concavité tournée vers le haut. La valeur minimale est associée à la position centrale (la moyenne des rangs). Les valeurs maximales correspondent aux individus situés aux extrémités (les rangs faibles et les rangs élevés). Les scores de Klotz, de Mood, de Siegel-Tuckey, de Savage, et de Ansari-Bradley appartiennent à cette catégorie.

5.4.1 Score de Wilcoxon

C'est la transformation la plus simple, $f(\cdot)$ correspond à la fonction identité, le score est égal au rang

$$s_i = f(r_i) = r_i \quad (5.5)$$

Ce score est utilisé dans le test de Wilcoxon-Mann-Whitney pour $K = 2$, dans le test de Kruskal-Wallis pour $K \geq 2$. Ils s'appliquent à la détection du décalage entre les paramètres de location des caractéristiques de tendance centrale des distributions conditionnelles.

Ce sont des tests *tout-terrain*, ils remplacent avantageusement le test de Student et l'ANOVA dès que l'on s'écarte de l'hypothèse de normalité des distributions (voir [3], page 154). Autre avantage non négligeable, ils sont très populaires, décrits dans quasiment tous les manuels de statistique. On y fait référence dès que l'on parle de tests non paramétriques de comparaison de populations. Il faut les connaître.

5.4.2 Score de la médiane

La fonction attribue le score 1 si le rang est supérieur à la médiane, 0 dans le cas contraire. Elle s'écrit

$$s_i = f(r_i) = \begin{cases} 1 & \text{si } r_i > \frac{n+1}{2} \\ 0 & \text{si } r_i \leq \frac{n+1}{2} \end{cases} \quad (5.6)$$

Le score de la médiane est bien entendu associé au test de la médiane pour $K = 2$ (test de Mood) et au test de la médiane généralisée pour $K \geq 2$ (test de Brown-Mood), utilisés pour mettre en évidence les décalages entre les caractéristiques de tendance centrale des distributions conditionnelles.

Ces tests sont indiqués lorsque la distribution des données sous-jacente est à queue lourde, ou lorsque les données sont tronquées car certaines d'entre elles excèdent une limite prédéfinie.

5.4.3 Score de Van der Waerden

Les tests associés mettent en évidence les écarts entre les paramètres de localisation. La fonction score correspond à l'inverse de la fonction de répartition de la loi normale centrée réduite appliquée sur les rangs. Elle s'écrit

$$s_i = f(r_i) = \Phi^{-1}\left(\frac{r_i}{n+1}\right) \quad (5.7)$$

Elle est particulièrement indiquée lorsque la distribution des données est proche de la loi normale.

Une fonction score quasi-équivalente est celle de Fisher-Yates-Terry-Hoeffding

$$s_i = f(r_i) = \Phi^{-1}\left(\frac{r_i - 0.375}{n + 0.25}\right) \quad (5.8)$$

Les tests qui en découlent ne se distinguent guère des tests de Van der Waerden en termes de comportement (puissance, etc.).

5.4.4 Score de Mood

Le score de Mood sert aux tests de comparaison des caractéristiques d'échelle des distributions. Il est égal au carré de l'écart de chaque rang avec la moyenne globale des rangs :

$$s_i = f(r_i) = \left[r_i - \frac{n+1}{2} \right]^2 \quad (5.9)$$

Les tests associés sont meilleurs que leur homologue paramétriques (test de Fisher et test de Bartlett) lorsque les données ne sont pas gaussiennes. Ils sont en revanche un peu moins efficace dans le cas contraire.

5.4.5 Score de Klotz

Le score de Klotz est égal au carré du score de Van der Waerden. Il permet de construire des tests qui mettent en évidence des différences dans les paramètres d'échelle des distributions. Sa formule est la suivante

$$s_i = f(r_i) = \left[\Phi^{-1}\left(\frac{r_i}{n+1}\right) \right]^2 \quad (5.10)$$

Les tests associés sont aussi efficaces que leurs homologues paramétriques lorsque les données sont gaussiennes, ils les surclassent dans les autres configurations. De manière générale, les tests issus de ce score sont meilleurs que les tests issus du score de Mood, à l'exception des distributions à queues lourdes (distribution de Cauchy par exemple).

5.4.6 Score de Savage

Elle est utilisée pour les tests de disparités d'échelle entre les distributions lorsque la distribution sous-jacente est exponentielle. Mais elle peut également servir à des tests de différences de paramètres de localisation lorsque les distributions sous-jacentes prennent une forme assez particulière (voir [3], pages 173 et 174).

$$s_i = f(r_i) = \sum_{j=1}^{r_i} \frac{1}{n-j+1} - 1 \quad (5.11)$$

5.4.7 Score de Siegel-Tukey

Il est utilisé pour le test du même nom destiné à comparer les paramètres d'échelle (voir [13], pages 156 à 160). La fonction est assez particulière, $f(r_i)$ résulte de la suite

$$f(1) = 1; f(n) = 2; f(n-1) = 3; f(2) = 4; f(3) = 5; f(n-2) = 6; f(n-3) = 7; \text{etc.}$$

jusqu'à ce que l'on parvienne à la moyenne des rangs.

5.4.8 Score de Ansari-Bradley

Elle est similaire au score de Siegel-Tukey, la fonction score maintenant résulte de la suite

$$f(1) = 1; f(n) = 1; f(2) = 2; f(n-1) = 2; f(3) = 3; f(n-3) = 3; \text{etc.}$$

Nous avons la possibilité de simplifier l'écriture

$$s_i = f(r_i) = \frac{n+1}{2} - \left| r_i - \frac{n+1}{2} \right| \quad (5.12)$$

L'idée est toujours la même : les scores élevés aux extrêmes, les scores faibles vers les valeurs centrales. L'efficacité des tests issus de cette fonction score est similaire à ceux issus du score de Siegel-Tukey (voir [3], page 173).

5.4.9 Traitement des ex-aequo

Les ex-aequo sont traités à travers **la méthode des scores moyens**. La procédure est la suivante :

- Les données sont triées selon la variable d'intérêt X croissante.
- Les rangs r_i sont attribués en numérotant simplement les observations selon leur position dans le tableau trié.
- Puis, à l'aide de la fonction $f(\cdot)$, le score $s_i = f(r_i)$ est calculé pour chaque individu.
- On recense les différentes valeurs distinctes G que prend X . Bien évidemment, $G \leq n$. Si $G = n$, il n'y a pas d'ex-aequo dans le fichier.
- Les valeurs rencontrées dans le fichier sont donc $\{v_1, v_2, \dots, v_g, \dots, v_G\}$.
- A chaque valeur v_g correspond t_g observations, nous calculons le score moyen

$$s'_g = \frac{1}{t_g} \sum_{i: x_i = v_g} s_i$$

- Ce score moyen sera alors affecté aux individus qui portent la valeur x_g .

Encore une fois, nous l'avions déjà précisé plus haut, il s'agit bien de la technique des scores moyens. La péréquation est réalisée après l'application de la fonction score. L'erreur à ne pas commettre serait d'appliquer la fonction score sur les rangs moyens. En effet, dans la très grande majorité des cas, la fonction est non linéaire.

5.4.10 Comparaison des fonctions score sur données simulées

A bien y regarder, les formulations génériques (section 5.2 et 5.3) correspondent à des tests de comparaison de scores moyens. Pour qu'elles traduisent une comparaison de paramètres de localisation ou d'échelle *sur les distributions initiales* des données, il faut définir judicieusement les fonctions scores.

Pour traduire cette idée, nous allons comparer le comportement du score de Van der Waerden et celui de Klotz dans la comparaison de $K = 2$ populations. Rappelons que ce dernier est simplement le carré du premier. Pourtant leur comportement n'est absolument pas le même :

1. La fonction de Van der Waerden est monotone selon le rang. Si les distributions sont décalées, les moyennes conditionnelles des scores seront différentes.
2. En revanche, la fonction de Klotz prend sa plus petite valeur lorsque nous nous rapprochons du rang moyen, elle prend des valeurs élevées dans le cas contraire. Elle est symétrique autour du rang moyen. De fait, *à condition que les paramètres de localisation conditionnels soient identiques c.-à-d. les rangs moyens conditionnels sont les mêmes dans les 2 sous populations*, si l'une des distributions est plus dispersée, elle aura tendance à prendre des valeurs de rangs à la fois faibles (queue de distribution à gauche) et fortes (queue de distribution à droite), son score moyen sera par conséquent plus élevé.

Dans ce qui suit, nous illustrons ces comportements à l'aide de données simulées.

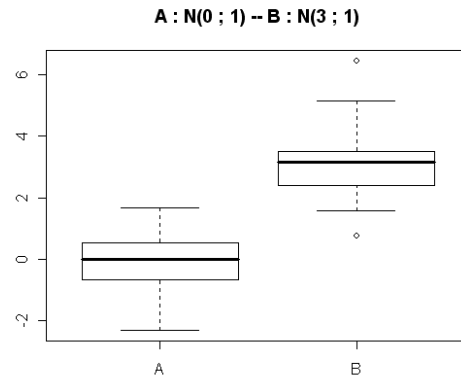


Fig. 5.1. Deux distributions gaussiennes - Décalage des paramètres de localisation

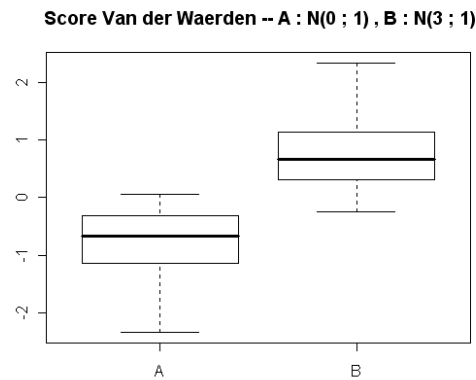


Fig. 5.2. Score de Van der Waerden sur $\mathcal{N}(0; 1)$ vs. $\mathcal{N}(3; 1)$

$\mathcal{N}(0; 1)$ vs. $\mathcal{N}(3; 1)$

Nous avons généré un échantillon de $n = 100$ observations. Il est composé de $n_1 = 50$ individus issus d'une loi $\mathcal{N}(0; 1)$ et $n_2 = 50$ autres d'une loi $\mathcal{N}(3; 1)$: deux sous échantillons gaussiens avec, clairement, des paramètres de localisation différents, et des dispersions identiques (Figure 5.1).

Nous appliquons la fonction score de Van der Waerden, nous visualisons maintenant les [distributions conditionnelles des scores](#) (Figure 5.2). Manifestement les moyennes des scores sont différentes. Le décalage entre les paramètres de localisation dans les distributions originales se traduisent par un décalage des scores moyens avec la fonction de Van der Waerden.

Voyons comment se comporte le score de Klotz sur les mêmes données (Figure 5.3). Cette fonction ne permet en rien de distinguer les différences des distributions conditionnelles originales. Les moyennes des scores sont confondues.

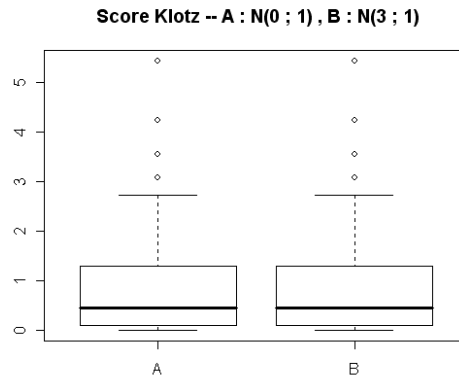


Fig. 5.3. Score de Klotz sur $\mathcal{N}(0;1)$ vs. $\mathcal{N}(3;1)$

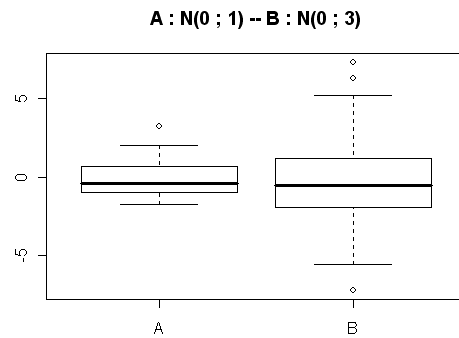


Fig. 5.4. Deux distributions gaussiennes - Différence des paramètres d'échelle

$\mathcal{N}(0;1)$ vs. $\mathcal{N}(0;3)$

Nous souhaitons maintenant comparer 2 sous-échantillons ($n_1 = n_2 = 50$) centrés mais de dispersions différentes (Figure 5.4).

Si nous appliquons la fonction score de Van der Waerden sur ces données. On se rend compte que la différence de dispersions dans les distributions conditionnelles originelles ne se traduit absolument pas en différence des moyennes de scores (Figure 5.5)⁵. Les tests basés sur les statistique de rang linéaires ne sont absolument pas opérants dans ce cas. La fonction n'est pas adaptée pour mettre en avant les différences d'échelle.

Lorsque nous passons au score de Klotz, la situation est tout autre (Figure 5.6). La différence de dispersion des distributions originelles sont traduites en différence des moyennes des scores. Le test de Klotz permettra de rejeter sans ambiguïté l'hypothèse nulle.

⁵ On remarquera incidemment que la transition par les rangs aura permis de museler les données atypiques (visibles dans les *boxplot* des distributions originelles - figure 5.4)

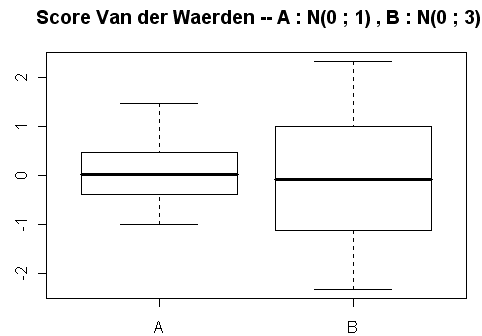


Fig. 5.5. Score de Van der Waerden sur $\mathcal{N}(0; 1)$ vs. $\mathcal{N}(0; 3)$

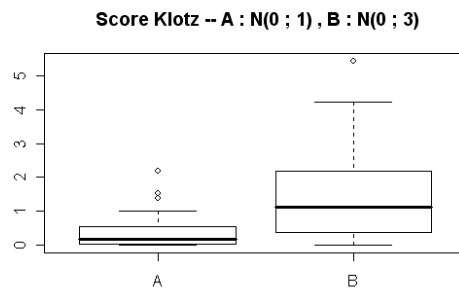


Fig. 5.6. Score de Klotz sur $\mathcal{N}(0; 1)$ vs. $\mathcal{N}(0; 3)$

Remarque 16 (Fonction score et distributions sous-jacente des données). Bien entendu, ces exemples n'ont pas valeur de preuve. Ils illustrent le rôle des fonctions score dans le mécanisme des tests basés sur les statistiques de rangs linéaires. Nous avons vu principalement leur impact sur les différences que l'on souhaite mettre en exergue. Une expérimentation tout aussi intéressante serait d'étudier leur comportement selon la distribution sous-jacente des données. Dans nos simulations, le choix des fonctions de Van der Waerden et de Klotz était particulièrement heureux dans la mesure où nos données ont été générées à partir de processus gaussiens.

5.5 Quelques exemples (repris des chapitres précédents)

A la lumière de cette nouvelle présentation standardisée des tests non paramétriques basés sur des statistique de rang linéaires, nous devrions pouvoir reprendre n'importe quel exemple décrit dans les chapitres précédents et les reproduire à l'aide des procédures génériques pour $K = 2$ et $K \geq 2$, où la seule modification porterait uniquement sur le choix de la fonction score. Voyons si les calculs concordent.

5.5.1 Test de Wilcoxon-Mann-Whitney

Reprenons l'exemple développé dans la section consacrée au test de Wilcoxon-Mann-Whitney (section 2.1.3, figure 2.4, page 40). Il s'agit de comparer l'indice de masse corporelle de jeunes adultes selon leur inclination au sport.

Nous sommes dans un cadre à $K = 2$ populations, nous utiliserons les formules de la section 5.2, avec la fonction score de Wilcoxon (section 5.4.1). Les calculs sont reproduits dans une feuille EXCEL (Figure 5.7) :

Wilcoxon - Mann - Whitney				
Imc	Sport	Score = Rang		
22.8	DAILY	1	n	14
23.4	DAILY	3	n1	7
23.6	DAILY	4	S = S1	39
23.7	DAILY	5		
24.8	DAILY	6	n2	7
26.1	DAILY	8	s_barre	7.5
30.2	DAILY	12		
23	NEVER	2	E0(S)	52.5
26	NEVER	7	V0(S)	61.25
26.3	NEVER	9		
27.3	NEVER	10	Z	-1.7250
28.7	NEVER	11		
33.5	NEVER	13		
35.3	NEVER	14		

Fig. 5.7. Imc selon activité sportive - Test de Mann-Whitney - Approche générique

- Le fichier comporte $n = 14$ individus, $n_1 = 7$ d'entre-eux font du sport quotidiennement, $n_2 = 7$ jamais.
- Nous rajoutons une colonne pour les rangs, elle correspond aussi à la colonne des scores dans le cas de Wilcoxon. Il n'y a pas d'ex-aequo dans ce fichier, aucun ajustement n'est nécessaire.
- Nous calculons la somme des rangs pour le 1^{er} groupe "Sport = Daily" qui sert de référence dans notre exemple, nous obtenons la statistique du test

$$S = S_1 = 1 + 3 + 4 + 5 + 6 + 8 + 12 = 39$$

- La moyenne globale des scores, calculée sur la totalité des observations est

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i = \frac{1}{14} (1 + 3 + \dots + 14) = 7.5$$

Ce résultat est évident puisque les scores sont égaux aux rangs dans le cas de la fonction de Wilcoxon, $\bar{s} = \bar{r} = \frac{n+1}{2} = \frac{14+1}{2} = 7.5$

- L'espérance mathématique de la statistique sous H_0 est égal à la moyenne globale des scores, multipliée par l'effectif n_1

$$E_0(S) = \frac{n_1}{n} \sum_{i=1}^n s_i = n_1 \times \bar{s} = 7 \times 7.5 = 52.5$$

- La variance sous H_0 est aussi calculée sur la globalité de l'échantillon

$$\begin{aligned}
V_0(S) &= \frac{n_1 n_2}{n(n-1)} \left[\sum_{i=1}^n (s_i - \bar{s})^2 \right] \\
&= \frac{7 \times 7}{14(14-1)} [(1-7.5)^2 + (2-7.5)^2 + \dots + (14-7.5)^2] \\
&= 61.25
\end{aligned}$$

- Nous pouvons maintenant construire la statistique centrée et réduite qui suit asymptotiquement une loi normale $\mathcal{N}(0;1)$

$$Z = \frac{S - E_0(S)}{\sqrt{V_0(S)}} = \frac{39 - 52.5}{\sqrt{61.25}} = -1.7250$$

- Nous retrouvons ainsi exactement le résultat calculé à l'aide de l'approche classique de la statistique du test de Wilcoxon-Mann-Whitney (Figure 2.4, sans la correction de continuité).

5.5.2 Test de Van der Waerden

Nous reprenons l'exemple dédié au test de Van der Waerden (section 17, figure 2.13, page 54). Il cumule 2 avantages : la fonction score n'est pas triviale, nous devons gérer les ex-aequo selon le principe des scores moyens. Nous sommes toujours dans un cadre à $K = 2$ populations.

L'objectif est de comparer le niveau d'anxiété chez deux populations, une ayant une tradition orale expliquant les effets de la maladie, l'autre non. La feuille de calcul est organisée selon toujours notre schéma générique (Figure 5.8) :

- Les données sont triées selon la variable d'intérêt "Anxiété".
- Nous assignons les rangs bruts r_i , puis les scores s_i .
- Comme il y a des ex-aequo, nous effectuons la péréquation selon la méthodes des scores moyens pour les individus portant des valeurs identiques, une nouvelle colonne s'_i est créée.
- A partir de ces informations, nous pouvons former la statistique de test. Nous avons $n = 39$, $n_1 = 16$ et $n_2 = 23$.
- La somme des scores pour le 1^{er} groupe qui nous sert de référence est

$$S = S_1 = -9.4126$$

- La moyenne des scores sur l'ensemble des observations est égale à

$$\bar{s} = \frac{1}{39}((-1.802) + (-1.802) + (-1.168) + \dots + 1.645 + 1.960) = 0.000$$

- L'espérance mathématique sous H_0 de S s'écrit

$$E_0(S) = n_1 \times \bar{s} = 16 \times 0.000 = 0.0000$$

- et la variance

Van der Waerden (avec gestion des ex-aequo)					
Tradition	Anxiete	Rang brut	F _i	s _i	s' _i
absent	6	1	0.025	-1.960	-1.802
present	6	2	0.050	-1.645	-1.802
absent	7	3	0.075	-1.440	-1.168
absent	7	4	0.100	-1.282	-1.168
absent	7	5	0.125	-1.150	-1.168
absent	7	6	0.150	-1.036	-1.168
absent	7	7	0.175	-0.935	-1.168
absent	8	8	0.200	-0.842	-0.717
absent	8	9	0.225	-0.755	-0.717
present	8	10	0.250	-0.674	-0.717
present	8	11	0.275	-0.598	-0.717
absent	9	12	0.300	-0.524	-0.524
absent	10	13	0.325	-0.454	-0.256
absent	10	14	0.350	-0.385	-0.256
absent	10	15	0.375	-0.319	-0.256
absent	10	16	0.400	-0.253	-0.256
present	10	17	0.425	-0.189	-0.256
present	10	18	0.450	-0.126	-0.256
present	10	19	0.475	-0.063	-0.256
present	11	20	0.500	0.000	0.031
present	11	21	0.525	0.063	0.031
absent	12	22	0.550	0.126	0.288
absent	12	23	0.575	0.189	0.288
present	12	24	0.600	0.253	0.288
present	12	25	0.625	0.319	0.288
present	12	26	0.650	0.385	0.288
present	12	27	0.675	0.454	0.288
absent	13	28	0.700	0.524	0.638
present	13	29	0.725	0.598	0.638
present	13	30	0.750	0.674	0.638
present	13	31	0.775	0.755	0.638
present	14	32	0.800	0.842	0.938
present	14	33	0.825	0.935	0.938
present	14	34	0.850	1.036	0.938
present	15	35	0.875	1.150	1.290
present	15	36	0.900	1.282	1.290
present	15	37	0.925	1.440	1.290
present	16	38	0.950	1.645	1.645
present	17	39	0.975	1.960	1.960

n	39
n1	16
S = S1	-9.4126
n2	23
s_barre	0.000
E0(S)	0.0000
V0(S)	8.0506
Z	-3.3174

Fig. 5.8. Analyse de l'anxiété - Test de Van der Waerden - Approche générique

$$\begin{aligned}
 V_0(S) &= \frac{n_1 n_2}{n(n-1)} \left[\sum_i (s'_i - \bar{s})^2 \right] \\
 &= \frac{16 \times 23}{39(39-1)} \times 32.4210 \\
 &= 8.0506
 \end{aligned}$$

– La valeur absolue de la statistique centrée réduite pour un test bilatéral est alors égal à

$$|Z| = \frac{|S - E_0(S)|}{\sqrt{V_0(S)}} = \frac{|-9.4126 - 0.0000|}{\sqrt{8.0506}} = 3.3174$$

– Soit exactement la valeur obtenue avec la méthode directe (figure 2.13).

Notons une information très importante dans ce schéma de calcul : une fois les scores assignés aux individus, intégrant le principe des scores moyens pour les individus présentant des valeurs identiques, **nous pouvons dérouler tous les calculs sans avoir à introduire explicitement une correction de la variance pour tenir compte des ex-aequo**. C'est une propriété de généralité remarquable. Moins nous aurons de cas particuliers à gérer, plus facile sera l'implémentation des techniques.

5.5.3 Test de Kruskal-Wallis

Nous reprenons dans cette section l'exemple des portées de porcs de Kruskal et Wallis (section 3.1.3, figure 3.4, page 73). Nous devons gérer 2 difficultés : $K > 2$ et traitement des ex-aequo. Nous utilisons la fonction score de Wilcoxon c.-à-d. les scores correspondent aux rangs simplement.

Nous mettons en oeuvre l'approche générique adéquate (section 5.3). L'organisation de la feuille Excel est la suivante (Figure 5.9) :

- Nous disposons de $n = 56$ observations.
- Nous trions le tableau de données selon la variable d'intérêt "Poids".
- Nous assignons les scores (rangs) aux individus, en corrigeant selon le principe des scores moyens en cas d'ex-aequo.
- Nous calculons la moyenne globale des scores

$$\bar{s} = 28.5$$

- De même, nous calculons la variance estimée des scores

$$T^2 = \frac{1}{n-1} \sum_i (s_i - \bar{s})^2 = 264.545$$

- Puis, pour chaque groupe, nous calculons les effectifs n_k et la somme des scores S_k . Ainsi, $n_1 = 10$, $n_2 = 8$, etc. , et $S_1 = 317$, $S_2 = 216.5$, etc.
- Toujours pour chaque groupe, nous calculons l'espérance mathématique des scores conditionnels sous H_0 c.-à-d. $E_0(S_k) = n_k \bar{s}$. Ainsi, $E_0(S_1) = 10 \times 28.5 = 285$, etc.
- Nous formons alors les $\frac{(S_k - E_0(S_k))^2}{n_k}$, dont la somme forme le numérateur de la statistique de test

$$Numerateur = 102.4 + 16.531 + 1664.1 + \dots + 1650 + 441 = 4911.396$$

- Le dénominateur correspond à T^2 .
- Au final, nous obtenons la statistique

$$C = \frac{Numerateur}{Denominateur} = \frac{4911.396}{264.545} = 18.565$$

- Ce qui correspond exactement à la statistique de test obtenue par la méthode directe (figure 3.4).

Ici également, **nous parvenons directement à la statistique corrigée sans avoir à introduire explicitement le facteur de correction pour les ex-aequo**. Notre seule contrainte est d'utiliser les scores moyens lors des calculs (rangs moyens dans le cas de Kruskal-Wallis). Ce résultat est valable quelle que soit la fonction score utilisée.

Kruskal-Wallis (avec gestion des ex-aequo)		
poide	portée	score = rang
1.1	c1	1
1.2	c7	2.5
1.2	c7	2.5
1.4	c8	4
1.6	c2	5
1.9	c1	6
2.0	c1	8.5
2.0	c2	8.5
2.0	c5	8.5
2.0	c5	8.5
2.1	c5	11
2.2	c7	12.5
2.2	c7	12.5
2.3	c2	14
2.4	c2	15.5
2.4	c8	15.5
2.5	c4	18.5
2.5	c6	18.5
2.5	c7	18.5
2.5	c8	18.5
2.6	c3	23
2.6	c4	23
2.6	c5	23
2.6	c5	23
2.6	c7	23
2.8	c1	27.5
2.8	c1	27.5
2.8	c2	27.5
2.8	c4	27.5
2.9	c3	31.5
2.9	c4	31.5
2.9	c5	31.5
2.9	c6	31.5
3.0	c8	34
3.1	c3	36
3.1	c6	36
3.1	c6	36
3.2	c1	41
3.2	c2	41
3.2	c3	41
3.2	c3	41
3.2	c3	41
3.2	c4	41
3.2	c4	41
3.3	c1	47.5
3.3	c1	47.5
3.3	c3	47.5
3.3	c3	47.5
3.3	c4	47.5
3.3	c4	47.5
3.4	c3	51
3.5	c2	52.5
3.5	c2	52.5
3.6	c1	54.5
3.6	c3	54.5
4.4	c1	56

n	56
s_barre	28.5
T²	264.545

portée	c1	c2	c3	c4	c5	c6	c7	c8
n_k	10	8	10	8	6	4	6	4
S_k	317	216.5	414	277.5	105.5	122	71.5	72

E0(S_k)	285	228	285	228	171	114	171	114
[S_k-E0(S_k)]²/n_k	102.4	16.531	1664.1	306.28	715.04	16	1650	441

Numérateur	4911.396
Dénominateur	264.545

C	18.565
---	--------

Fig. 5.9. Analyse des poids à la naissance - Test de Kruskal-Wallis - Approche générique

Tests pour échantillons appariés

Tests pour ($K = 2$) échantillons liés

6.1 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 hasard¹ 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 prétendus 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 appellations (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>.

6.2 Test des signes

6.2.1 Test d'hypothèses, statistique de test et région critique

Définition du test

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 $D = X_1 - X_2$ dont les valeurs d_i sont obtenues par différences des paires de valeurs c.-à-d.

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

Le test des signes consiste à s'intéresser uniquement au sens de l'écart $+$ ($D > 0$) ou $-$ ($D < 0$), et non à son importance. Il est adapté à tout type de variables (quantitative, ordinale, même binaire) dès lors qu'il est possible de déterminer si une valeur est plus importante qu'une autre pour chaque paire d'observation. Ce gain en champ d'application entraîne en contrepartie une perte de puissance du test si d'aventure l'amplitude de l'écart était une information exploitable dans l'étude. Son homologue paramétrique est le test de Student pour échantillons appariés².

Soit $\pi = \Pr(D > 0)$, le test d'hypothèses, s'il est bilatéral s'écrit

$$\begin{aligned} H_0 : \pi &= \frac{1}{2} \\ H_1 : \pi &\neq \frac{1}{2} \end{aligned}$$

Pour un test unilatéral, nous pouvons former l'hypothèse alternative $H_1 : \pi > \frac{1}{2}$ (resp. $H_1 : \pi < \frac{1}{2}$) lorsqu'on souhaite savoir si X_1 est stochastiquement plus grand (resp. plus petit) que X_2 .

² Voir <http://udel.edu/~mcdonald/statsign.html>

Statistique de test et région critique

Schématiquement, le test des signes consiste simplement à tester si le nombre des écarts positifs est significativement différent du nombre des écarts négatifs (plus élevée ou plus rare concernant les tests unilatéraux).

Pour un échantillon de taille n , une statistique naturelle pour le test des signes est le nombre d'occurrence S des écarts positifs c.-à-d.

$$S = \sum_{i=1}^n \delta_i \quad (6.1)$$

où $\delta_i = 1$ si $d_i > 0$, $\delta_i = 0$ si $d_i < 0$.

Sous H_0 , S suit une loi binomiale $\mathcal{B}(n, \frac{1}{2})$. Comme cette dernière est symétrique, il est possible d'obtenir directement la probabilité critique (p-value) du test bilatéral en calculant la quantité

$$p = 2 \times \Pr[\mathcal{B}(n, \frac{1}{2}) \geq \max(S, n - S)] \quad (6.2)$$

Pour un test unilatéral $H_1 : \pi > \frac{1}{2}$, la probabilité critique sera $p = \Pr[\mathcal{B}(n, \frac{1}{2}) \geq \max(S, n - S)]$.

Pour rappel, la probabilité cumulée d'une loi binomiale s'écrit

$$\begin{aligned} P[\mathcal{B}(n, \pi) \geq S] &= \sum_{j=S}^n \binom{n}{j} \pi^j (1 - \pi)^{n-j} \\ &= \sum_{j=S}^n \frac{n!}{j!(n-j)!} \pi^j (1 - \pi)^{n-j} \end{aligned}$$

Exemple 32 (Effet du glucose sur la mémoire). Nous reprenons cet exemple de [6], page 739. Il s'agit d'identifier l'effet du glucose sur la mémoire de $n = 16$ patients âgés. L'expérimentation mise en place est la suivante : au réveil, on leur donne une boisson sucrée au moyen de glucose, on leur narre une histoire. Quelque temps après, on leur demande de la retracer. La qualité de la restitution est notée par un juge. Après un délai raisonnable, on recommence la même expérimentation mais en leur donnant une boisson à la saccharine (Remarque : on peut aussi imaginer que l'ordre d'administration des sucreries est défini aléatoirement selon les sujets). L'objectif de l'étude est de vérifier que l'ingestion de la boisson au glucose entraîne une qualité de mémorisation différente.

Les données sont recensées dans une feuille Excel (Figure 6.1) :

- Nous retrouvons les colonnes de valeurs pour X_1 (glucose) et X_2 (saccharine). Nous disposons de $n = 16$ observations.
- Nous formons une nouvelle colonne qui prend la valeur "+" ou "-" selon le signe de $d_i = x_{i1} - x_{i2}$.
- Nous pouvons former la statistique $S = \sum_i \delta_i = 16$ correspondant au nombre de signe "+".

X1 (glucose)	X2 (saccharine)	Signe(écart)
0	1	--
10	9	+
9	6	+
4	2	+
8	5	+
6	5	+
9	7	+
3	2	+
12	8	+
10	8	+
15	11	+
9	3	+
5	6	--
6	8	--
10	8	+
6	4	+

n	16
S	13

j	Pr
13	0.00854
14	0.00183
15	0.00024
16	0.00002
Somme	0.01064

p-value	0.02127
---------	---------

Fig. 6.1. Effet du glucose sur la mémorisation - Test des signes

– Pour calculer la quantité $\Pr[\mathcal{B}(16, \frac{1}{2}) \geq 13]$, nous énumérons les cas :

$$\Pr[\mathcal{B}(16, \frac{1}{2}) \geq 13] = \binom{16}{13} (0.5)^{13} (1 - 0.5)^{16-13} = 0.00854$$

$$\Pr[\mathcal{B}(16, \frac{1}{2}) \geq 14] = 0.00183$$

$$\Pr[\mathcal{B}(16, \frac{1}{2}) \geq 15] = 0.00024$$

$$\Pr[\mathcal{B}(16, \frac{1}{2}) \geq 16] = 0.00002$$

La somme fait $\Pr[\mathcal{B}(16, \frac{1}{2}) \geq 13] = 0.01064$

- Dans le cadre d'un test bilatéral, la probabilité critique du test est égal à $p = 2 \times 0.01064 = 0.02127$. Au risque 5%, nous pouvons rejeter l'hypothèse nulle d'égalité des effets du glucose et de la saccharine sur la qualité de la mémorisation.
- Pour un test unilatéral $H_1 : \pi > \frac{1}{2}$, la probabilité critique serait directement $p = \Pr[\mathcal{B}(16, \frac{1}{2}) \geq 13] = 0.01064$. On conclurait alors que l'effet du glucose sur la mémorisation est significativement meilleur.

Le cas particulier des écarts nuls

Lorsqu'il y a des écarts nuls dans les données, c.-à-d. $d_i = x_{i1} - x_{i2} = 0$, la modélisation des signes via la loi binomiale pose problème. En effet, il y aurait à ce moment là, 3 résultats possibles ($d_i < 0; d_i = 0; d_i > 0$). Plutôt que de s'embarquer dans des schémas compliqués, on conseille généralement de supprimer ces observations (voir [3], page 186 ; [6], pages 736 et 737 ; [13], page 83). On réduit d'autant l'effectif n dans les calculs. Voyons ce mécanisme sur un exemple.

Exemple 33 (La prise de décision dans un couple). Cet exemple est tiré de [13], page 81. On a demandé à des couples d'évaluer le poids de leur conjoint dans la prise de décision lors des achats importants. L'homme note l'influence de sa femme, et vice versa. L'étude porte sur 17 couples. Les données sont regroupés dans une feuille Excel (Figure 6.2), on cherche à savoir si l'homme donne plus d'importance à sa femme qu'inversement, nous sommes dans le cadre d'un test unilatéral :

Numéro	Mari	Femme	Ecart	Signe
A	5	3	2	+
B	4	3	1	+
C	6	4	2	+
D	6	5	1	+
E	3	3	0	0
F	2	3	-1	--
G	5	2	3	+
H	3	3	0	0
I	1	2	-1	--
J	4	3	1	+
K	5	2	3	+
L	4	2	2	+
M	4	5	-1	--
N	7	2	5	+
O	5	5	0	0
P	5	3	2	+
Q	5	1	4	+

Nombre de couples	17
$n (+ \text{ ou } -)$	14
S	11

j	Pr
11	0.02222
12	0.00555
13	0.00085
14	0.00006
Somme	0.02869

p-value	0.02869
---------	---------

Fig. 6.2. Influence des époux lors de la décision d'achat - Test des signes

- Le nombre de couples dans l'étude est 17.
- Mais lorsque nous comptabilisons les écarts, nous nous rendons compte qu'il est nul (égal à 0) pour les couples $\{E, H, O\}$. On décide donc de les exclure des calculs. Nous disposons finalement de $n = 14$ observations exploitables.
- Nous comptons les observations telles que $d_i > 0$, nous obtenons $S = 11$.
- Nous calculons alors la probabilité $\Pr[\mathcal{B}(14, \frac{1}{2}) \geq 11] = 0.02869$
- Qui est égal à la probabilité critique du test aussi puisque nous réalisons un test unilatéral.
- Pour un niveau de signification de 5%, les hommes pensent que leur épouse ont une place plus importante dans la prise de décision lors des achats importants.

6.2.2 L'approximation normale pour les grands effectifs

Pour n assez grand, ($n > 35$) en pratique (voir [13], page 83), la distribution de S , qui est binomiale, peut être approchée par la loi normale qui, sous $H_0 : \pi = \frac{1}{2}$, prend les paramètres

$$E(S) = n \times \pi = \frac{n}{2} \quad (6.3)$$

$$V(S) = n \times \pi \times (1 - \pi) = \frac{n}{4} \quad (6.4)$$

La statistique centrée et réduite du test s'écrit

$$Z = \frac{S - n/2}{\sqrt{n/4}} = \frac{2S - n}{\sqrt{n}} \quad (6.5)$$

Pour un test bilatéral au risque α , la région critique du test est définie par

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

Remarque 17 (Correction de continuité). Pour les effectifs modérés, nous pouvons améliorer la précision en introduisant une correction de continuité. Pour le test bilatéral, il s'écrit

$$|Z| = \frac{|2S - n| - 1.0}{\sqrt{n}}$$

Pour les tests unilatéraux, nous utiliserons

$$Z = \frac{2S - n \pm 1.0}{\sqrt{n}}$$

Nous rajoutons +1 pour un test unilatéral à gauche ($H_1 : \pi < \frac{1}{2}$); nous retranchons pour un test unilatéral à droite.

Exemple 34 (Shooté à l'hélium). On a demandé à 39 footballeurs de shooter dans des ballons gonflés à l'hélium et à l'air. On cherche à savoir si dans le premier cas, ils arrivent à envoyer le ballon plus loin³. Nous sommes dans le cadre d'un test unilatéral à droite $H_1 : \pi > \frac{1}{2}$. La feuille Excel se présente comme suit (Figure 6.3) :

- Nous disposons de 39 observations initialement. Après avoir calculé les différences entre X_1 (hélium) et X_2 (air), nous écartons les observations $n^o 1$ et $n^o 21$ car $d_i = 0$ pour ces cas. Nous disposons donc de $n = 37$ observations exploitables.
- Nous calculons le nombre de différences positives $S = 20$.
- Nous pouvons alors produire la statistique centrée et réduite, avec la correction de continuité -1 puisque nous sommes dans le cadre d'un test unilatéral à droite

$$Z = \frac{(2S - n) - 1}{\sqrt{n}} = \frac{(2 \times 20 - 37) - 1}{\sqrt{37}} = 0.32880$$

- La probabilité critique associée est $p = 0.37115$. Pour un niveau de signification 5%, nous ne pouvons pas rejeter l'hypothèse nulle : gonfler les ballons à l'hélium ne permet pas de les envoyer plus loin.

6.2.3 Sorties des logiciels

Pour illustrer les productions des logiciels, nous utilisons le fichier "Crédit". Nous cherchons à savoir si, à l'intérieur des couples, l'homme perçoit un salaire plus élevé que sa femme. Nous sommes bien en présence d'échantillons appariés. Chaque couple représente un bloc. Nous disposons initialement de 50 couples.

TANAGRA affiche les moyennes et écarts type des variables. A titre indicatif, car le calcul n'est absolument pas basé sur ces quantités. Puis il comptabilise le nombre d'observations pour lesquelles l'écart est non nul "Used examples" (Figure 6.4). Dans cet exemple, nous utiliserons finalement $n = 49$ observations pour les calculs. Le nombre d'écarts positifs "Homme - Femme > 0 " est $S = 33$. TANAGRA

³ Pour une description plus complète des conditions de recueil des données, voir <http://lib.stat.cmu.edu/DASL/Stories/Heliumfootball.html>

Essai	Helium	Air	Ecart	Signe
1	25	25	0	0
2	16	23	-7	--
3	25	18	7	+
4	14	16	-2	--
5	23	35	-12	--
6	29	15	14	+
7	25	26	-1	--
8	26	24	2	+
9	22	24	-2	--
10	26	28	-2	--
11	12	25	-13	--
12	28	19	9	+
13	28	27	1	+
14	31	25	6	+
15	22	34	-12	--
16	29	26	3	+
17	23	20	3	+
18	26	22	4	+
19	35	33	2	+
20	24	29	-5	--
21	31	31	0	0
22	34	27	7	+
23	39	22	17	+
24	32	29	3	+
25	14	28	-14	--
26	28	29	-1	--
27	30	22	8	+
28	27	31	-4	--
29	33	25	8	+
30	11	20	-9	--
31	26	27	-1	--
32	32	26	6	+
33	30	28	2	+
34	29	32	-3	--
35	30	28	2	+
36	29	25	4	+
37	29	31	-2	--
38	30	28	2	+
39	26	28	-2	--

Essais 39

n (+ ou -) 37

S 20

Z 0.32680

p-value 0.37115

Fig. 6.3. Gonflage à l'hélium de ballon - Distance de shoot - Test des signes

Results					
Attribute_Y		Attribute_X		Statistical test	
Sal.Homme		Sal.Femme		Measure	Value
Avg	7.464000	Avg	7.309400	Used examples	49
Std-dev	0.561863	Std-dev	0.509948	Positive	33
				Negative	16
				Z	2.285714
				Pr(> Z)	0.022271

Fig. 6.4. Fichier "Crédit", écart des salaires homme-femme - TANAGRA

utilise directement l'approximation normale. Il construit la statistique centrée et réduite pour un test bilatéral. La correction de continuité est introduite, soit

$$|Z| = \frac{|2S - n| - 1.0}{\sqrt{n}} = \frac{|2 \times 33 - 49| - 1.0}{\sqrt{49}} = 2.285714$$

La probabilité critique basée sur l'approximation normale est donc $p = 0.022271$.

R, à travers le **package BSDA**, introduit d'autres informations très intéressantes dans ses résultats (Figure 6.5). La statistique $S = 33$ est fournie. Il utilise en revanche la *distribution binomiale* pour calculer la probabilité critique du test, soit

$$p = 2 \times \sum_{j=33}^{n=49} \binom{n}{j} \left(\frac{1}{2}\right)^n = 0.02129$$

Pour $n = 49$, l'écart est faible par rapport à l'approximation normale.

R affiche de surcroît une estimation de la médiane de l'écart "Salaire.Homme - Salaire.Femme" $\hat{m} = 0.2$, et son intervalle de confiance au niveau 95% (option modifiable), soit $[0.0307; 0.2939]$ en ce qui concerne nos données.

```

R Console
> sign.test(donnees$Sal.Homme, donnees$Sal.Femme)
$rval

Dependent-samples Sign-Test

data:  donnees$Sal.Homme and donnees$Sal.Femme
S = 33, p-value = 0.02129
alternative hypothesis: true median difference is not equal to 0
95 percent confidence interval:
 0.03070229 0.29394656
sample estimates:
median of x-y
      0.2

$Confidence.Intervals
              Conf.Level L.E.pt U.E.pt
Lower Achieved CI    0.9351 0.0400 0.2800
Interpolated CI      0.9500 0.0307 0.2939
Upper Achieved CI    0.9672 0.0200 0.3100

Warning message:
In return(rval, Confidence.Intervals) :
  les renvois multi-arguments sont obsolètes
>

```

Fig. 6.5. Fichier "Crédit", écart des salaires homme-femme - R

6.3 Test des rangs signés de Wilcoxon

6.3.1 Test d'hypothèses, statistique de test et région critique

Le test des rangs signés de Wilcoxon traite la comparaison d'échantillons appariés. Il répond donc à la même catégorie de problèmes que le test des signes⁴. Nous ne devons pas le confondre avec le test des

⁴ Voir <http://www.cons-dev.org/elearning/stat/St4.html> ; <http://udel.edu/~mcdonald/statsignedrank.html>

rangs de Wilcoxon-Mann-Whitney pour échantillons indépendants, même si le mécanisme du test repose sur une somme de rangs.

Par rapport au test des signes, le test de Wilcoxon **utilise l'importance relative des écarts** lors de la définition de la statistique de test. Il est donc plus riche, plus puissant pour peu que l'on puisse les ordonner (les écarts).

Par rapport au test de comparaison de moyennes (test de Student) pour échantillons appariés, il n'exploite pas l'amplitude de l'écart. Cet inconvénient se révèle être un avantage dans certaines situations : (1) le test ne suppose pas la normalité sous-jacente des données, son champ d'application est donc plus large, comme tous les tests non paramétriques ; (2) l'effet dévastateur des points atypiques est considérablement amoindri, canalisant le rôle de quelques observations qui peuvent totalement fausser les calculs⁵.

Bref, le test de rangs signés de Wilcoxon se comporte très bien face au test équivalent paramétrique (test de Student sur échantillons appariés). Il fait face lors que les conditions d'application du test de Student sont réunies, il le surclasse dès qu'on s'en éloigne (voir [13], page 95 ; [3], pages 194 à 198, avec notamment des résultats sur des données simulées).

Construction de la statistique de test

De nouveau, nous travaillons sur les écarts $d_i = x_{i1} - x_{i2}$, mais nous les exploitons de manière différente :

1. Nous construisons la valeur absolue des écarts $|d_i|$
2. **A partir de $|d_i|$, la valeur absolue des écarts, nous définissons les rangs des observations r_i .** Ainsi la plus petite valeur des écarts, en valeur absolue, reçoit le rang 1 ; le plus grand écart, toujours en valeur absolue, se voit attribué le rang n .
3. A partir des rangs, nous pouvons calculer la statistique T^+ , elle correspond à la somme des rangs pour les observations présentant un écart positif ($d_i > 0$)

$$T^+ = \sum_{i:d_i>0} r_i \quad (6.6)$$

4. De la même manière, nous pourrions définir T^- pour les écarts négatifs. Mais sachant que la somme totale des rangs est égale à $\frac{n(n+1)}{2}$, nous pouvons déduire T^- à partir de la relation suivante

$$T^- = \frac{n(n+1)}{2} - T^+$$

⁵ Si un des écarts prend une valeur absolue 100 fois plus forte que les autres, en travaillant sur les rangs, nous ne tenons compte que de l'information "c'est le plus grand écart". Il n'écrase donc pas les autres observations lors de la formation de la statistique de test.

Sous H_0 , les deux variables X_1 et X_2 ont une fonction de répartition identique, ou plus précisément, leurs paramètres de localisation sont équivalents, nous aurons l'égalité $T^+ = T^- = \frac{1}{2} \left[\frac{n(n+1)}{2} \right]$

Plus T^+ sera grand (resp. petit) par rapport à T^- , plus nous serons amenés à conclure que les valeurs prises par X_1 sont stochastiquement plus élevés (resp. plus faibles) par rapport à celles de X_2 .

Pour un test bilatéral, la région critique du test au risque α s'écrit

$$R.C. : (T^+ \leq \frac{n(n+1)}{2} - T_\alpha) \quad \text{ou} \quad (T^+ \geq T_\alpha)$$

où le seuil critique T_α est lue dans une table spécifique due a Wilcoxon (section D.5; voir aussi <http://www.cons-dev.org/elearning/stat/Tables/Tab5.html>)

Exemple 35 (Effets de l'entraînement sur la tension artérielle). On cherche à savoir si un entraînement régulier modifie la tension artérielle des personnes (test bilatéral). On a recueilli la tension systolique de $n = 8$ personnes, on leur a fait suivre un programme d'entraînement spécifique pendant 6 mois, puis on leur a de nouveau mesuré la tension (voir [6], page 735).

N°	Tension		Ecart	Ecart	Rang(ecart)	Rang signé	
	Avant	Après					
1	130	120	10	10	5	5	n = 8
2	170	163	7	7	4	4	
3	125	120	5	5	2	2	T+ = 27
4	170	135	35	35	7	7	
5	130	143	-13	13	6	-6	
6	130	136	-6	6	3	-3	
7	145	144	1	1	1	1	
8	160	120	40	40	8	8	

Fig. 6.6. Évolution de la tension artérielle après entraînement - Test des rangs signés

Les données et les calculs sont réunies dans une feuille Excel (Figure 6.6) :

- Nous disposons de $n = 8$ paires d'observations.
- Nous calculons les écarts entre les valeurs "Avant" et "Après", c'est le rôle de la colonne "Écarts"
- Puis nous déduisons la valeur absolue de l'écart $|d_i|$ qui va servir de base de calcul pour les rangs.
- La colonne "Rangs" ne pose pas de problèmes particulier. Par exemple, pour l'individu $n^o 5$, $d_5 = -13$, et $r_5 = \text{rang}(|d_5|) = 6$.
- Nous rajoutons une colonne supplémentaire, les "rangs signés" par commodité de calculs. Nous nous en servons pour additionner toutes valeurs positives pour obtenir la statistique T^+

$$T^+ = 5 + 4 + 2 + 7 + 1 + 8 = 27$$

- Dans la table des seuils critique, pour un test bilatéral à 5%, nous lisons $T_{0.05} = 4$ pour un échantillon de taille $n = 8$. Manifestement, nous sommes dans la région critique. La tension artérielle n'est plus la même après un entraînement de 6 mois.

Traitement des écarts nuls

De nouveau ici, à l'instar du test des signes, les écarts nuls $d_i = 0$ posent problèmes car ils ne rentrent pas dans le canevas de notre modélisation. **La solution usuelle consiste tout simplement à supprimer les observations correspondantes.**

De fait, la "vraie" taille d'échantillon n correspond au nombre d'observations pour lesquelles $d_i \neq 0$. Les rangs seront calculés uniquement sur ces individus.

Le principe des rangs moyens pour les ex-aequo

La situation est différente lorsque nous avons des ex-aequo. Dans ce cas, plusieurs observations présentent une valeur identique de $|d_i|$, nous devons leur attribuer un rang identique.

Nous utilisons le principe des rangs moyens pour traiter cette de configuration. Le rapprochement avec la démarche mise en place lors de l'analyse d'échantillons indépendants est approprié (page 36). La statistique de test T^+ n'est pas modifiée. Sa variance en revanche le sera. Ce qui aura un impact lorsque nous passons à l'approximation normale pour les grands échantillons, et qu'il faudra utiliser la statistique centrée et réduite Z pour définir la région critique.

Pour l'heure, illustrons le principe des rangs moyens sur un exemple pour bien comprendre le mécanisme.

Exemple 36 (Influence des époux lors de la décision d'achat). Nous reprenons l'exemple des époux qui jugent leurs influences respectives lors des importantes décision d'achat (ce test a déjà été utilisé pour illustrer le test des signes, figure 6.2). Nous souhaitons illustrer 2 configurations : comment exclure les observations à écarts nuls $d_i = 0$ des calculs ; comment mettre en place le principe des rangs moyens pour les ex-aequo.

N°	Mari	Femme	d_i	$ d_i $	r_i	r'_i	rang signé
E	3	3	0	0	-	-	-
H	3	3	0	0	-	-	-
O	5	5	0	0	-	-	-
B	4	3	1	1	1	3.5	3.5
D	6	5	1	1	2	3.5	3.5
F	2	3	-1	1	3	3.5	-3.5
I	1	2	-1	1	4	3.5	-3.5
J	4	3	1	1	5	3.5	3.5
M	4	5	-1	1	6	3.5	-3.5
A	5	3	2	2	7	8.5	8.5
C	6	4	2	2	8	8.5	8.5
L	4	2	2	2	9	8.5	8.5
P	5	3	2	2	10	8.5	8.5
G	5	2	3	3	11	11.5	11.5
K	5	2	3	3	12	11.5	11.5
Q	5	1	4	4	13	13	13
N	7	2	5	5	14	14	14

n

14

T+

94.5

Fig. 6.7. Influence des époux lors de la décision d'achat - Test des rangs signés

La feuille Excel retrace les opérations (Figure 6.7) :

- Au départ, nous disposons de 17 observations.
- A partir des variables originelles "Mari" et "Femme", nous calculons la différence d_i , puis la valeur absolue de la différence $|d_i|$.
- Nous trions alors le tableau selon $|d_i|$ croissant. Nous excluons les observations pour lesquelles $|d_i| = 0$ c.-à-d. les époux $\{E, H, O\}$.
- Sur les $n = 14$ observations restantes, nous calculons les rangs bruts r_i . Comme le tableau est déjà trié, il suffit d'attribuer dans l'ordre les numéros 1, 2, ...
- Nous remarquons néanmoins que plusieurs observations ont des valeurs identiques de $|d_i|$, par exemple les époux $\{B, D, F, I, J, M\}$ présentent tous $|d_i| = 1$, nous devons leur attribuer un rang identique.
- Pour ce faire, nous créons une nouvelle colonne des rangs moyens r'_i , pour tous les groupes d'observations ayant une valeur de $|d_i|$ identiques, nous effectuons une péréquation des rangs. Voyons ce qu'il en est pour le premier groupe correspondant à $|d_i| = 1$ avec 6 observations, le rang moyen associé sera

$$r' = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

Ainsi, les époux $\{B, D, F, I, J, M\}$ se voient tous attribuer le même rang $r' = 3.5$.

Nous faisons de même pour tous les groupes d'observations.

- Nous déduisons alors la colonne des rangs signés à partir des r'_i .
- La statistique de test T^+ est obtenue en faisant la somme des rangs signés positifs (en fond vert dans la feuille Excel), soit

$$T^+ = 3.5 + 3.5 + 3.5 + 8.5 + 8.5 + 8.5 + 8.5 + 11.5 + 11.5 + 13 + 14 = 94.5$$

- Pour un test bilatéral à 5%, le seuil critique du test pour $n = 14$ est égal à $T_{0.05} = 21$.
- Nous sommes dans la région critique, très largement même, manifestement hommes et femmes n'évaluent pas de la même manière leurs influences respectives lors des décisions d'achat.
- En creusant un peu les résultats, on se rend compte que lorsque l'écart est faible ($|d_i| = 1$), la situation est à peu près symétrique; mais dès qu'il devient important ($|d_i| \geq 2$), les hommes attribuent systématiquement (sur ces données en tous les cas) un poids plus important à leur épouse qu'inversement lors de la décision d'achat.

6.3.2 Approximation normale pour les grands effectifs

Approximation normale

Lorsque n est assez grand, ($n > 15$) en pratique (voir [13], page 15), on peut approximer la distribution de T^+ sous H_0 par une loi normale de paramètres

$$E(T^+) = \frac{1}{4}n(n+1) \quad (6.7)$$

$$V(T^+) = \frac{1}{24}n(n+1)(2n+1) \quad (6.8)$$

La statistique de test centrée réduite suit une loi normale, son expression est la suivante

$$Z = \frac{T^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \quad (6.9)$$

La région critique pour un test bilatéral au risque α s'écrit

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

Remarque 18 (Correction de continuité). Bien sûr, pour les effectifs modérés, nous pouvons améliorer la précision à l'aide de la correction de continuité. Nous utiliserons pour le test bilatéral

$$|Z| = \frac{|T^+ - \frac{1}{4}n(n+1)| - 0.5}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$$

Correction de la variance pour les ex-aequo

Lorsque les $|d_i|$ comportent des ex-aequo, nous avons vu plus haut comment mettre en place le mécanisme des rangs moyens. Corollaire à cet ajustement, nous devons également corriger la variance de la statistique de test.

Soit G le nombre de valeurs distinctes de $|d_i|$, pour la valeur $n^\circ g$, nous observons t_g observations, la variance corrigée s'écrit alors (voir [3], page 191)

$$\tilde{V}(T^+) = \frac{1}{24}n(n+1)(2n+1) - \frac{1}{48} \sum_{g=1}^G t_g(t_g-1)(t_g+1) \quad (6.10)$$

La statistique centrée réduite destinée à définir la région critique est également modifiée

$$\tilde{Z} = \frac{T^+ - \frac{1}{4}n(n+1)}{\sqrt{\tilde{V}(T^+)}}$$

Exemple 37 (Les ballons gonflés à l'hélium). Nous reprenons l'exemple "Shooté à l'hélium" étudié lors de la présentation de l'approximation normale de la statistique du test des signes (page 127). On cherche à savoir si les shoots vont plus loin lorsque l'on gonfle des ballons avec de l'hélium plutôt qu'avec de l'air. Voyons ce que nous propose le test des rangs signés⁶ (Figure 6.8) :

- Le tableau comporte initialement 39 observations, décrites par les variables "Hélium" et "Air".
- Nous calculons les écarts d_i et la valeur absolue $|d_i|$. Nous trions le tableau selon $|d_i|$ croissant.
- Nous constatons qu'il y a 2 écarts nuls dans ce fichier, ils correspondent aux observations $n^\circ 1$ et $n^\circ 21$. Ils sont exclus du mouvement.
- Nous disposons en réalité de $n = 37$ observations exploitables.

⁶ Pour éviter la confusion, nous n'introduirons pas la correction de continuité dans le calcul des statistiques centrées et réduites.

Essai	Helium	Air	d_i	$ d_i $	r_i	r'_i	rang signé
1	25	25	0	0	-	-	-
21	31	31	0	0	-	-	-
7	25	26	-1	1	1	2.5	-2.5
13	28	27	1	1	2	2.5	2.5
26	28	29	-1	1	3	2.5	-2.5
31	26	27	-1	1	4	2.5	-2.5
4	14	16	-2	2	5	9.5	-9.5
8	26	24	2	2	6	9.5	9.5
9	22	24	-2	2	7	9.5	-9.5
10	26	28	-2	2	8	9.5	-9.5
19	35	33	2	2	9	9.5	9.5
33	30	28	2	2	10	9.5	9.5
35	30	28	2	2	11	9.5	9.5
37	29	31	-2	2	12	9.5	-9.5
38	30	28	2	2	13	9.5	9.5
39	26	28	-2	2	14	9.5	-9.5
16	29	26	3	3	15	16.5	16.5
17	23	20	3	3	16	16.5	16.5
24	32	29	3	3	17	16.5	16.5
34	29	32	-3	3	18	16.5	-16.5
18	26	22	4	4	19	20	20
28	27	31	-4	4	20	20	-20
36	29	25	4	4	21	20	20
20	24	29	-5	5	22	22	-22
14	31	25	6	6	23	23.5	23.5
32	32	26	6	6	24	23.5	23.5
2	16	23	-7	7	25	26	-26
3	25	18	7	7	26	26	26
22	34	27	7	7	27	26	26
27	30	22	8	8	28	28.5	28.5
29	33	25	8	8	29	28.5	28.5
12	28	19	9	9	30	30.5	30.5
30	11	20	-9	9	31	30.5	-30.5
5	23	35	-12	12	32	32.5	-32.5
15	22	34	-12	12	33	32.5	-32.5
11	12	25	-13	13	34	34	-34
6	29	15	14	14	35	35.5	35.5
25	14	28	-14	14	36	35.5	-35.5
23	39	22	17	17	37	37	37

n	37
T+	398.5
E(T+)	351.5

Sans correction pour ex-aequo

V(T+)	4393.75
Z	0.70906
p-value	0.23914

Avec correction pour ex-aequo

Tableau des $ d_i $ distinctes et comptage			
g	valeur $ d_i $	t _g	t _g (t _g -1)(t _g +1)
1	1	4	60
2	2	10	990
3	3	4	60
4	4	3	24
5	5	1	0
6	6	2	6
7	7	3	24
8	8	2	6
9	9	2	6
10	12	2	6
11	13	1	0
12	14	2	6
13	17	1	0
Somme			1188

$\sim V(T+)$	4369
$\sim Z$	0.71106
p-value	0.23852

Fig. 6.8. Gonflage à l'hélium de ballon - Test des rangs signés de Wilcoxon

- Nous attribuons les rangs bruts r_i à partir de $|d_i|$ (tel que $|d_i| > 0$).
- Il y a de nombreux ex-aequo sur les valeurs de $|d_i|$. Nous mettons en oeuvre le principe des rangs moyens, une nouvelle colonne r'_i est produite. Prenons un exemple pour illustrer le mécanisme. Les individus $n^o 7, 13, 26, 31$ partagent la valeur $|d_i| = 1$. Nous faisons la péréquation des rangs de manière à leur attribuer le rang moyen

$$r' = \frac{1 + 2 + 3 + 4}{4} = 2.5$$

Etc.

- La colonne des rangs signés permet de retrouver le signe des écarts sur les rangs.
- La statistique du test T^+ correspond à la somme des rangs positifs, soit

$$T^+ = 2.5 + 9.5 + \dots + 35.5 + 37 = 398.5$$

- L'espérance mathématique de T^+ sous H_0 s'obtient par

$$E(T^+) = \frac{1}{4}n(n+1) = \frac{1}{4}37(37+1) = 351.5$$

- Dans un premier temps, nous nous intéressons à la *variance non corrigée des ex-aequo*. Nous obtenons,

$$V(T^+) = \frac{1}{24}n(n+1)(2n+1) = \frac{1}{24}37(37+1)(2 \times 37+1) = 4393.75$$

- La statistique centrée réduite Z est égale à

$$Z = \frac{|T^+ - \frac{1}{4}n(n+1)| - 0.5}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} = \frac{398.5 - 351.3}{\sqrt{4393.75}} = 0.70906$$

- Pour un risque unilatéral à 5%, nous devons comparer cette valeur à $u_{0.95} = 1.645$. Puis $Z = 0.70906 < u_{0.95} = 1.645$, nous ne pouvons pas rejeter l'hypothèse nulle.
- La probabilité critique du test est $p = 0.23914$

Il y a un assez grand nombre d'ex-aequo sur les valeurs $|d_i|$, en toute rigueur la variance doit être corrigée. Voyons l'impact de cette correction sur la statistique centrée et réduite du test :

- Tout d'abord, nous devons comptabiliser les valeurs distinctes de $|d_i|$. Ils sont au nombre de $G = 13$. Le tableau "Tableau des $|d_i|$ distinctes et comptage" énumère ces valeurs et les effectifs associés t_g .
- Dans la dernière colonne de ce tableau, nous pouvons alors la quantité $t_g(t_g - 1)(t_g + 1)$ et en faire la somme, nous obtenons

$$\sum_{g=1}^G t_g(t_g - 1)(t_g + 1) = 1188$$

- Nous pouvons ainsi calculer la variance ajustée

$$\begin{aligned}\tilde{V}(T^+) &= V(T^+) - \frac{1}{48} \sum_{g=1}^G t_g(t_g - 1)(t_g + 1) \\ &= 4393.75 - \frac{1}{48} \times 1188 \\ &= 4369\end{aligned}$$

- La statistique centrée et réduite du test revient maintenant à

$$\tilde{Z} = \frac{398.5 - 351.5}{\sqrt{4369}} = 0.71106$$

- La probabilité critique du test unilatéral est $p = 0.23852$. Définitivement, gonfler les ballons à l'hélium n'ont jamais aidé les pieds carrés à jouer aux Platini.
- Remarquons que s'il n'y pas d'ex-aequo dans les $|d_i|$ c.-à-d. $t_g = 1, \forall g$, la correction est nulle $\tilde{V}(T^+) = V(T^+)$.
- De manière générale, $\tilde{V}(T^+) \leq V(T^+)$, la statistique centrée et réduite est toujours (au moins) augmentée lorsque l'on introduit la correction pour ex-aequo.

6.3.3 Sorties des logiciels

Nous reprenons la comparaison des salaires masculins et féminins dans le fichier "Crédit". Le test des signes avait détecté un écart significatif (page 128). Or nous savons qu'il est conservateur car il ne tient

Results					
Attribute_Y		Attribute_X		Statistical test	
Sal.Homme		Sal.Femme		Measure	Value
Avg	7.464000	Avg	7.309400	Used examples	49
Std-dev	0.561863	Std-dev	0.509948	Sum ranks + (T+)	954.5
				Sum ranks - (T-)	270.5
				E(T+)	612.500000
				V(T+)	10102.625000
				Z	3.402585
				Pr(> Z)	0.000668

Fig. 6.9. Écart des salaires homme-femme - Test des rangs signés - TANAGRA

pas compte que du sens des écarts et non de leur importance relative. Voyons ce qu'il en est avec le test des rangs signés, nous devrions obtenir des résultats plus tranchés encore.

TANAGRA produit à titre indicatif les statistique descriptives conditionnelles, essentiellement la moyenne et l'écart type. Plus intéressant pour notre test est la colonne "Statistical test". Nous y observons (Figure 6.9) :

- Le nombre d'observations pour lesquels l'écart est non nul, nous avons $n = 49$. Elles seront utilisées pour la construction de la statistique de test.
- La somme des rangs positifs est $T^+ = 954.5$, celle des rangs négatifs est $T^- = 270.5$. Si la théorie est vraie, nous devrions avoir $T^+ + T^- = \frac{n(n+1)}{2}$, ce qui est le cas, $954.5 + 270.5 = \frac{49(49+1)}{2} = 1225$. Ce genre de petites vérifications ne fait jamais de mal.
- Sous H_0 , l'espérance de la statistique de test est

$$E(T^+) = \frac{1}{4}n(n+1) = \frac{1}{4}49(49+1) = 612.5$$

sa variance, tenant compte des ex-aequo (équation 6.10)

$$\tilde{V}(T^+) = 10102.625$$

- La statistique centrée et réduite est obtenue à l'aide de

$$Z = \frac{954.5 - 612.5}{\sqrt{10102.625}} = 3.402585$$

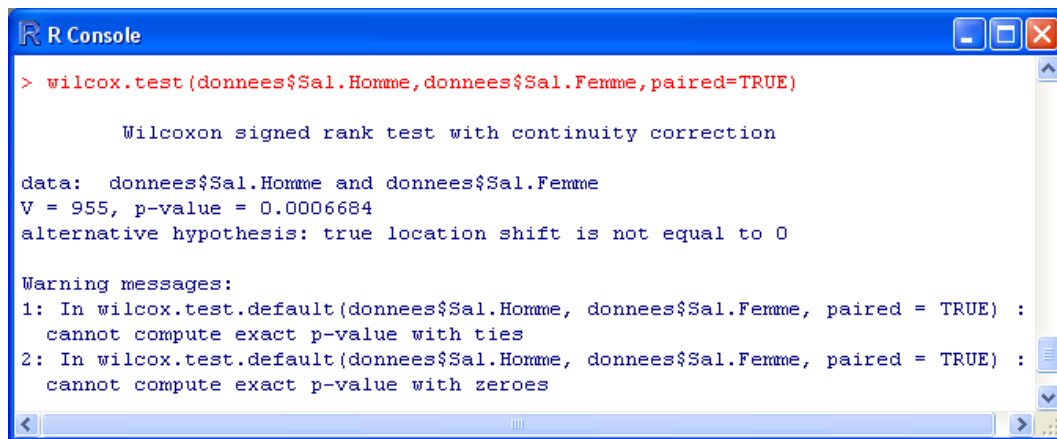
Avec un probabilité critique, pour un test bilatéral, $p = 0.000668$.

Remarque 19 (Et si on veut introduire la correction de continuité ?). Nous avons la possibilité de mettre en place des variantes. Si nous souhaitons introduire la correction de continuité pour un test bilatéral, il suffit de faire, à partir des informations proposées dans le tableau "Statistical Test"

$$|Z| = \frac{|954.5 - 612.5| - \frac{1}{2}}{\sqrt{10102.625}} = 3.397611$$

La probabilité critique sera alors $p = 0.000680$.

R, toujours aussi lapidaire, produit des résultats très proches de ceux de TANAGRA. Il semble que la procédure implémentée ne tienne compte ni des ex-aequo, ni des écarts nuls. Toujours est-il que R nous propose $T^+ = 955$, avec une probabilité critique $p = 0.0006684$ (Figure 6.10).



```
R Console
> wilcox.test(donnees$Sal.Homme,donnees$Sal.Femme,paired=TRUE)

      Wilcoxon signed rank test with continuity correction

data:  donnees$Sal.Homme and donnees$Sal.Femme
V = 955, p-value = 0.0006684
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(donnees$Sal.Homme, donnees$Sal.Femme, paired = TRUE) :
   cannot compute exact p-value with ties
2: In wilcox.test.default(donnees$Sal.Homme, donnees$Sal.Femme, paired = TRUE) :
   cannot compute exact p-value with zeroes
```

Fig. 6.10. Écart des salaires homme-femme - Test des rangs signés - Logiciel R

Tests pour ($K \geq 2$) échantillons liés

7.1 Appariement pour $K \geq 2$ échantillons - Les blocs aléatoires complets

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 [4], chapitre 6, pages 141 à 167).

En anglais, le terme consacré est *randomized blocks*¹. Reprenons notre exemple des additifs pour carburants. Nous souhaitons maintenant comparer $K = 5$ marques différentes. De la même manière 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, les blocs doivent être indépendants les uns des autres, et il ne doit pas y avoir d'interaction blocs-traitements.

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".

¹ Voir <http://www.socialresearchmethods.net/kb/expblock.php> pour une description détaillée de la stratégie et des bénéfices qu'on peut en attendre

7.2 ANOVA de Friedman

7.2.1 Principe, statistique de test et région critique

L'analyse de variance (ANOVA) de Friedman consiste à comparer K paramètres de localisation sur K échantillons liés. Le tableau de données comporte donc n lignes, et K colonnes (pour nous ce sont des variables, on parle aussi de "traitements" dans la terminologie des plans d'expériences). Dans sa construction, chaque ligne correspond à un "K-uplet" de mesures. Ces dernières peuvent être continues ou ordinales, l'essentiel est que l'on puisse les exploiter de manière à produire un classement c.-à-d. affecter des rangs aux traitements. Par hypothèse, les formes des distributions des K échantillons sont identiques, quand bien même elles seraient décalées, en particulier elles doivent présenter une dispersion identique.

Le test de Friedman est une généralisation du test des signes (voir [3], page 241).

Le test d'hypothèses s'écrit

$$\begin{aligned} H_0 : \theta_1 &= \dots = \theta_K \\ H_1 : \exists k, k' \quad \text{tel que} \quad \theta_k &\neq \theta_{k'} \end{aligned}$$

Il s'agit bien d'un test de comparaison de K paramètres de localisation. Mais, à la différence du test de Kruskal-Wallis par exemple, les valeurs seront uniquement comparables à l'intérieur de chaque groupe. Ainsi, **nous travaillerons bien sur des rangs, mais calculés à l'intérieur des blocs**. C'est ce qui différencie le test de Friedman de tous les tests pour échantillons indépendants où les rangs étaient toujours construits à partir de la totalité de observations mesurées.

Notre tableau des rangs qui servira au calcul de la statistique de test se présente comme suit

Bloc vs. Traitement	X_1	\dots	X_k	\dots	X_K	Somme
1	r_{11}	\dots	r_{1k}	\dots	r_{1K}	$R_{1.} = K(K+1)/2$
\dots						\dots
i	r_{i1}	\dots	r_{ik}	\dots	r_{iK}	$R_{i.} = K(K+1)/2$
\dots						\dots
n	r_{n1}	\dots	r_{nk}	\dots	r_{nK}	$R_{n.} = K(K+1)/2$
Somme	$S_{.1}$	\dots	$S_{.k}$	\dots	$S_{.K}$	—
Moyenne	$\bar{R}_{.1}$	\dots	$\bar{R}_{.k} = S_{.k}/n$	\dots	$\bar{R}_{.K}$	—

Première information importante, les rangs étant calculés à l'intérieur de chaque bloc, la somme des rangs d'un bloc est toujours égal à $R_{i.} = \frac{K(K+1)}{2}$. Analyser cette quantité n'a aucun sens.

Pour comparer l'efficacité des traitements, nous devons donc nous intéresser à la somme des rangs par traitements $S_{.k}$. Ainsi, la statistique du test de Friedman s'écrit

$$Fr = \frac{12n}{K(K+1)} \sum_{k=1}^K (\bar{R}_{.k} - \bar{\bar{R}})^2 \quad (7.1)$$

où $\bar{\bar{R}}$ est la moyenne globale des rangs, soit

$$\bar{\bar{R}} = \frac{\sum_i \sum_k r_{ik}}{nK} = \frac{nK(K+1)/2}{nK} = (K+1)/2$$

L'écriture ci-dessus correspond alors à

$$Fr = \frac{12n}{K(K+1)} \sum_{k=1}^K (\bar{R}_{.k} - \frac{n+1}{2})^2$$

En simplifiant un peu, nous retrouvons une des écritures usuelles de la statistique de test ([3], page 242)

$$Fr = \frac{12n}{K(K+1)} \sum_{k=1}^K \bar{R}_{.k}^2 - 3n(K+1) \quad (7.2)$$

L'autre écriture courante consiste à exprimer la statistique à partir de la somme des rangs ([13], page 176; [6], page 743)

$$Fr = \frac{12}{nK(K+1)} \sum_{k=1}^K S_{.k}^2 - 3n(K+1) \quad (7.3)$$

Détaillons ces formules pour comprendre la nature du test. $\bar{R}_{.k}$ est un indicateur de performance du traitement $n^{\circ}k$. Si $\bar{R}_{.k} = 1$, cela veut dire que le traitement est systématiquement le moins bon dans tous les blocs; à l'inverse, si $\bar{R}_{.k} = K$, il est systématiquement le meilleur. De fait, avec la formulation 7.1, on comprend que Fr traduit l'idée d'une variabilité inter-traitements, elle correspond à la dispersion des rangs moyens conditionnels autour de la moyenne globale. Si les traitements se valent tous, nous obtiendrons $Fr = 0$. Plus ils se démarqueront les uns des autres, plus Fr prendra une valeur élevée.

On peut construire de manière différente la statistique Fr . On montre facilement que Fr est construit comme un rapport de variance inter-traitement et de variance intra-traitements sur les rangs².

La région critique du test correspond donc aux grandes valeurs de Fr , soit, au risque α :

$$R.C. : Fr \geq Fr_{\alpha,k,n}$$

où $Fr_{\alpha,k,n}$ est lue dans la table des valeurs critiques de la statistique de Friedman (section D.6)³.

Remarque 20 (Anova de Friedman, généralisation du test des signes). Le parallèle entre le test des signes et le test de Friedman est possible. Il a été montré en effet pour $K = 2$ que (voir [3], page 242)

$$Fr = \frac{4}{n} (S - \frac{n}{2})^2$$

où S est la statistique du test des signes. De ce point de vue, on peut considérer que le test de Friedman est une généralisation du test des signes qui s'applique pour la comparaison de $K \geq 2$ populations.

² Voir http://en.wikipedia.org/wiki/Friedman_test

³ D'autres tables sont disponibles en ligne : <http://courses.wcupa.edu/rbove/eco252/252suppkey.htm>; ou <http://www.cons-dev.org/elearning/ando/tables/tables/Table10.html>

Exemple 38 (Comparaison de marques de pneumatiques). On souhaite comparer la longévité de 4 marques de pneumatiques (traitement). Nous leur faisons parcourir 10 scénarios de parcours routier (blocs). La variable d'intérêt est la distance parcourue jusqu'à une certaine limite d'usure.

		Pneumatiques (Traitement)			
		A	F	G	R
Blocs	1	38	29	41.5	39
	2	24.5	36	35.5	25
	3	37.5	38.5	31.5	29.5
	4	20.5	33.5	29.5	21.5
	5	29.5	35	34	11
	6	22	33	35	17
	7	29	37.5	38	18.5
	8	25	21.5	28	18
	9	26	35.5	28	17
	10	17	23.5	34	16.5

Rangs				
A	F	G	R	
2	1	4	3	
1	4	3	2	
3	4	2	1	
1	4	3	2	
2	4	3	1	
2	3	4	1	
2	3	4	1	
3	2	4	1	
2	4	3	1	
2	3	4	1	
S _k	20	32	34	14
R _{barre_k}	2.0	3.2	3.4	1.4

S ² _k	400	1024	1156	196
-----------------------------	-----	------	------	-----

n	10
K	4

T	2776
---	------

Fr	16.56
----	-------

Seuil critique exact - Friedman	
Fr _{0.05,10,4}	7.68
Fr _{0.01,10,4}	10.68

Approximation KHI-2	
ddl	3
KHI2 _{0.99(3)}	11.34
p-value	0.001

Fig. 7.1. Longévité des pneumatiques - Test de Friedman

Les données et les calculs sont résumés dans une feuille Excel (Figure 7.1) :

- Nous disposons de $n = 10$ observations (blocs), et nous souhaitons comparer $K = 4$ marques de pneumatiques (traitements).
- Première étape indispensable, nous devons transformer les données. Les rangs sont calculés à l'intérieur des blocs. Ainsi, dans chacune des lignes du tableau "Rangs", nous devons observer les valeurs allant de 1 à $K = 4$.
- Pour évaluer les performances des marques de pneumatiques, nous formons les sommes des rangs en colonnes, nous obtenons les $S_{.k}$. Par exemple, pour la première marque, nous avons $S_{.1} = 20$; pour la seconde, $S_{.2} = 32$, etc.
- Nous pouvons également exprimer les performances en termes de moyenne des rangs $\bar{R}_{.k} = S_{.k}/n$. Ainsi, $R_{.1} = S_{.1}/n = 20/10 = 2.0$, etc.
- Pour calculer la statistique du test selon l'expression 7.3, nous calculons les quantités $S_{.k}^2$: $S_{.1}^2 = 400$, $S_{.2}^2 = 1024$, $S_{.3}^2 = 1156$, $S_{.4}^2 = 196$. La somme de ces quantités est égale à

$$T = \sum_{k=1}^4 S_{.k}^2 = 400 + 1024 + 1156 + 196 = 2776$$

- Nous pouvons maintenant produire la statistique du test

$$Fr = \frac{12}{nK(K+1)} \sum_{k=1}^K S_{.k}^2 - 3n(K+1) = \frac{12}{10 \times 4 \times (4+1)} \times 2776 - 3 \times 10 \times (4+1) = 16.56$$

- Intéressons-nous tout d'abord à l'approche exacte (encadré en vert continu). Le seuil critique du test à 5% est égal à 7.68 (voir table statistique en section D.6). Les marques de pneumatiques ne présentent pas des performances identiques au risque 5%.
- Pour un risque à 1%, le seuil est à 10.68 dans la table, nous pouvons en déduire que la probabilité critique p du test est < 0.01 .

Ces mêmes données ont été analysées à l'aide d'une comparaison de K moyennes pour les plans d'expérience en blocs aléatoires complets dans notre support consacré aux tests paramétriques⁴. On avait conclu à une différence de performances très significative entre les marques, avec une probabilité critique de $p = 0.00005$. Le résultat est cohérent avec le test de Friedman.

7.2.2 Approximations pour les grands échantillons

Lorsque n ou K sont *assez grands*, en pratique ($n > 15$ ou $K > 4$)⁵, la distribution de Fr peut être approximée par une loi du χ^2 à $(K-1)$ degrés de liberté. La région critique sera ainsi

$$R.C. : Fr \geq \chi_{1-\alpha}^2(K-1)$$

où $\chi_{1-\alpha}^2(K-1)$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $(K-1)$ degrés de liberté.

Exemple 39 (Approximation χ^2 - Tester la longévité des pneumatiques). Revenons sur notre exemple ci-dessus "Longévité des pneumatiques". Nous sommes à la lisière des conditions d'une bonne approximation avec $n = 10$ et $K = 4$. Essayons néanmoins, ne serait-ce que pour comparer les conclusions du test. Revenons sur notre feuille de calcul (Figure 7.1, la partie encadrée en pointillés bleu).

Le nombre de degrés de liberté est $ddl = K - 1 = 4 - 1 = 3$. Le seuil critique au risque 1% est le quantile $\chi_{0.99}^2(3) = 11.34$. Nous sommes dans la région critique $Fr = 16.56 > 11.34 = \chi_{0.99}^2(3)$. Les résultats restent cohérents avec l'approche exacte. La probabilité critique associée à la loi asymptotique est $p = 0.001$.

Notons quand même qu'il faut être extrêmement prudent sur d'aussi petits effectifs, il vaut mieux favoriser l'approche exacte tant que les tables sont disponibles. Dans notre exemple, la différence entre le seuil exact (10.68 à 1%) et le seuil asymptotique (11.34 au même niveau de risque) n'est pas négligeable, la décision peut être inversée lorsque nous sommes à la lisière de la région critique.

⁴ Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf

⁵ Voir http://en.wikipedia.org/wiki/Friedman_test; $n \times k \geq 30$ dans [3], page 243.

7.2.3 Traitement des ex-aequo

Lorsqu'il y a des *ex-aequo* à l'intérieur d'un bloc, le principe des rangs moyens est utilisé c.-à-d. on réalise la péréquation des rangs affectés aux traitements qui présentent la même valeur. La statistique de test doit être corrigée pour tenir compte de l'ajustement, avec les caractéristiques suivantes : la statistique corrigée est toujours supérieure ou égale à la statistique non corrigée ; s'il n'y a pas aucun ex-aequo, la statistique corrigée et non corrigée doivent être identiques ; la correction sera d'autant plus sensible que le nombre d'ex-aequo est élevé.

Pour le bloc $n^o i$: nous notons G_i le nombre de valeurs différentes, la valeur $n^o g$ étant répétée t_{ig} fois. La statistique de Friedman ajustée pour les ex-aequo s'écrit alors

$$\tilde{Fr} = \frac{Fr}{1 - \frac{\sum_{i=1}^n \sum_{g=1}^{G_i} (t_{ig}^3 - t_{ig})}{n(K^3 - K)}} \quad (7.4)$$

La quantité au dénominateur est toujours ≤ 1 , on comprend aisément que $\tilde{Fr} \geq Fr$. Sous H_0 , les lois asymptotiques de \tilde{Fr} et Fr sont identiques.

Exemple 40 (Torches de soudage). Six ($n = 6$) soudeurs avec différents niveaux de savoir faire ont été invités à souder deux tuyaux à l'aide cinq ($K = 5$) différentes torches de soudage. Les torches ont été affectés dans un ordre aléatoire pour chaque soudeur. La qualité de soudure a été évaluée par un juge (note allant de 1 à 10, 10 correspondant à la meilleure évaluation). On cherche à savoir si la qualité de soudure dépend du type de torche utilisé⁶.

Ces données nous intéressent surtout parce qu'elles comportent des ex-aequo dans certaines lignes du tableau. Nous utilisons les rangs moyens dans ce cas, et nous devons également corriger la statistique de test (Figure 7.2) :

- Dans une première étape, le tableau de données est transformé en tableau de rangs, toujours calculés ligne par ligne. Prenons le cas de la première ligne, nous observons les valeurs (après les avoir triés) $\{3.3, 3.9, 4.1, 4.1, 4.2\}$. Il y a un ex-aequo pour la valeur 4.1 qui occupent les 3^{eme} et 4^{eme} rangs. En vertu de la règle des rangs moyens, nous obtenons les rangs $\{1, 2, 3.5, 3.5, 5\}$.
- Nous appliquons le même principe aux autres lignes pour compléter le tableau des rangs.
- A partir de ce tableau, nous calculons les sommes de rangs $S_{.k}$ et les rangs moyens $\bar{R}_{.k}$ par colonne. Puis nous formons les quantités $S_{.k}^2$.
- Nous calculons alors la somme T

$$T = \sum_{k=1}^5 S_{.k}^2 = 552.25 + 484 + 420.25 + 324 + 36 = 1816.5$$

- Nous pouvons produire la statistique non corrigée de Friedman

$$Fr = \frac{12}{nK(K+1)} \sum_{k=1}^K S_{.k}^2 - 3n(K+1) = \frac{12}{6 \times 5 \times (5+1)} \times 1816.5 - 3 \times 6 \times (5+1) = 13.1000$$

⁶ Source <http://www.texasoft.com/winkfrie.html>

Bloc	Soudeur	Traitement				
		Torche1	Torche2	Torche3	Torche4	Torche5
	1	3.9	4.1	4.2	4.1	3.3
	2	9.4	9.5	9.4	9.0	8.6
	3	9.7	9.3	9.3	9.2	8.4
	4	8.3	8.0	7.9	8.6	7.4
	5	9.8	8.9	9.0	9.0	8.3
	6	9.9	10.0	9.7	9.6	9.1

	Rangs				
	Torche1	Torche2	Torche3	Torche4	Torche5
1	2	3.5	5	3.5	1
2	3.5	5	3.5	2	1
3	5	3.5	3.5	2	1
4	4	3	2	5	1
5	5	2	3.5	3.5	1
6	4	5	3	2	1
S_k	23.5	22	20.5	18	6
R_barre_k	3.92	3.67	3.42	3.00	1.00
S²_k	552.25	484	420.25	324	36
n	6				
K	5				
T	1816.5				
Fr	13.1000				

Calcul de (t _{ig} ³ - t _{ig})	
1	6
2	6
3	6
4	0
5	6
6	0
Total	24

Correction (C)	0.9667
~Fr	13.5517
ddl	4
kH12_0.95	9.4877
p-value	0.0089

Fig. 7.2. Influence des torches sur la qualité de soudure - Test de Friedman

- Occupons nous maintenant du facteur de correction au dénominateur de l'équation 7.4.
- Dans le sous-tableau "Calcul de ($t_{ig}^3 - t_{ig}$)", nous regroupons le calcul de la double somme $\sum_{i=1}^n \sum_{g=1}^{G_i} (t_{ig}^3 - t_{ig})$. Détaillons le calcul pour le 1^{er} bloc, nous listons les différentes valeurs et leur nombre d'apparition entre parenthèses : {3.3(1), 3.9(1), 4.1(2), 4.2(1)}. Nous pouvons former

$$\sum_{g=1}^4 (t_{1g}^3 - t_{1g}) = (1^3 - 1) + (1^3 - 1) + (2^3 - 2) + (1^3 - 1) = 0 + 0 + 6 + 0 = 6$$

- Nous faisons de même pour chaque bloc, la double somme est alors égale à

$$\sum_{i=1}^n \sum_{g=1}^{G_i} (t_{ig}^3 - t_{ig}) = 6 + 6 + 6 + 0 + 6 + 0 = 24$$

- Le facteur de correction est

$$C = 1 - \frac{\sum_{i=1}^n \sum_{g=1}^{G_i} (t_{ig}^3 - t_{ig})}{n(K^3 - K)} = 1 - \frac{24}{6 \times (5^3 - 5)} = 0.9667$$

- Et la statistique corrigée

$$\tilde{Fr} = \frac{Fr}{C} = \frac{13.1000}{0.9667} = 13.5517$$

- Puisque $K = 5 > 4$, nous pouvons utiliser la loi asymptotique, le seuil critique au risque 5% est $\chi_{0.95}^2(4) = 9.4877$. Nous nous situons dans la régions critique, $\tilde{Fr} = 13.5517 > 9.4877 = \chi_{0.95}^2(4)$. Nous rejetons l'hypothèse nulle d'influence égale des torches sur la qualité de la soudure.
- La probabilité critique du test est $p = 0.0089$, renforçant cette conclusion.

7.2.4 Sorties des logiciels

Nous reprenons l'exemple des "Torches de soudure" ci-dessus (Figure 7.2) pour commenter les sorties des logiciels. Il sera plus facile de suivre les résultats.

RANKS			Friedman Statistic	
Att.	Sum(Ranks)	Mean(Ranks)	Stat.	Value
Torche1	23.5	3.9167	Friedman Fr	13.55172
Torche2	22.0	3.6667	d.f.	4
Torche3	20.5	3.4167	p-value	0.00887
Torche4	18.0	3.0000		
Torche5	6.0	1.0000		

Fig. 7.3. Influence des torches sur la qualité de soudure - Test de Friedman avec TANAGRA

TANAGRA affiche les sommes et les moyennes des rangs pour chaque type de torche. A partir de ces valeurs, il est facile de reconstituer la statistique de test à l'aide de l'équation 7.3). **TANAGRA** fournit directement la statistique, éventuellement corrigée pour les ex-aequo, avec $Fr = 13.55172$. Le nombre de degré de liberté est $d.f = K - 1 = 5 - 1 = 4$, la probabilité critique $p = 0.00887$ (Figure 7.3). **R** fournit exactement les mêmes résultats (Figure 7.4).

```

R Console
> friedman.test(as.matrix(donnees))

Friedman rank sum test

data: as.matrix(donnees)
Friedman chi-squared = 13.5517, df = 4, p-value = 0.008872
> |

```

Fig. 7.4. Influence des torches sur la qualité de soudure - Test de Friedman avec R

7.3 Détermination de la source des écarts

Le rejet de l'hypothèse nulle lors du test de Friedman indique qu'un des paramètres de localisation au moins s'écarte des autres (ou d'une autre). Pour détecter les écarts significatifs, nous pouvons réaliser plusieurs types de comparaisons, soit en les testant toutes 2 à 2, soit en confrontant les $(K-1)$ traitements avec l'un d'entre eux qui sert de référence. Les tests d'hypothèses et les idées sous-jacentes ont été décrits dans la section 3.2 consacrée aux échantillons indépendants (page 75). Pour plus de détails, le mieux est de s'y référer. Les schémas sont les mêmes, mis à part que nous travaillons sur des échantillons appariés. Nous nous contenterons de donner directement les principales formules dans cette section.

7.3.1 Comparaisons multiples

Il s'agit de comparer toutes les catégories 2 à 2, à l'aide de tests bilatéraux du type

$$\begin{aligned} H_0 : \theta_j &= \theta_l \\ H_1 : \theta_j &\neq \theta_l \end{aligned}$$

Nous avons $K(K-1)/2$ comparaisons à réaliser. Pour conserver le risque global du test de Friedman, nous devons corriger le risque des comparaisons individuelles. Pour n assez grand, la région critique des tests s'écrit

$$|S_j - S_l| \geq u_{1-\alpha} \sqrt{\frac{nK(K+1)}{6}} \quad (7.5)$$

avec

$$\alpha = \frac{\alpha/2}{K(K-1)/2} = \frac{\alpha}{K(K-1)}$$

Remarque 21 (Définir les régions critiques à partir des rangs moyens). Nous pouvons également baser la comparaison sur les rangs moyens par catégorie, la région critique s'écrit alors

$$|\bar{R}_{.j} - \bar{R}_{.l}| \geq u_{1-\alpha} \sqrt{\frac{K(K+1)}{6n}} \quad (7.6)$$

Exemple 41 (Performances des pneumatiques). Revenons sur notre exemple concernant la longévité des pneumatiques (page 144). Le test de Friedman indiquait un écart de performances significatif pour au moins une des marques. Déterminons maintenant celles qui sont à la source des différences (Figure 7.5) :

- Le tableau de données avec $n = 10$ et $K = 4$ est transformé en tableau de rangs, puis nous calculons les sommes de rangs par colonnes S_k qui vont nous servir pour les comparaisons.

		Pneumatiques (Traitement)				Rangs			
		A	F	G	R	A	F	G	R
Blocs	1	38	29	41.5	39	2	1	4	3
	2	24.5	36	35.5	25	1	4	3	2
	3	37.5	38.5	31.5	29.5	3	4	2	1
	4	20.5	33.5	29.5	21.5	1	4	3	2
	5	29.5	35	34	11	2	4	3	1
	6	22	33	35	17	2	3	4	1
	7	29	37.5	38	18.5	2	3	4	1
	8	25	21.5	28	18	3	2	4	1
	9	26	35.5	28	17	2	4	3	1
	10	17	23.5	34	16.5	2	3	4	1
		S_k				20	32	34	14

n	10
K	4

alpha	0.05
a	0.00417

u_{1-a}	2.6383
Seuil	15.232

Tableau des écarts		
Comparaisons	Ecart	Signif.
A vs. F	12	non
A vs. G	14	non
A vs. R	6	non
F vs. G	2	non
F vs. R	18	oui
G vs. R	20	oui

Fig. 7.5. Comparaisons multiples à l'issue du Test de Friedman - Longévité des pneumatiques

- Le risque global choisi est $\alpha = 0.05$, nous en déduisons

$$a = \frac{\alpha}{K(K-1)} = \frac{0.05}{4 \times (4-1)} = 0.00417$$

- Le quantile de la loi normale utilisé pour les tests est

$$u_{1-a} = u_{1-0.00417} = 2.6383$$

- Le seuil critique des comparaisons sera donc

$$seuil = u_{1-a} \sqrt{\frac{nK(K+1)}{6}} = 2.6383 \sqrt{\frac{10 \times 4 \times (4+1)}{6}} = 15.232$$

- Pour toutes les combinaisons (A vs. F, A vs. G, ..., G vs. R), nous calculons les écarts absolus $|S_j - S_l|$, et nous les comparons au seuil. C'est le rôle du "Tableau des écarts".
- On se rend compte ainsi que les écarts significatifs concernent l'opposition "marque F vs. marque R", et "marque G vs. marque R".

7.3.2 Comparaisons à une référence

Tout comme pour les comparaisons consécutifs au test de kruskal-Wallis pour échantillons indépendants (section 3.2), il est possible de définir une catégorie (traitement) de référence. L'objectif est de déterminer celles qui s'en démarquent significativement, en conservant le niveau de risque global du test. Soit Ω_c la sous-population de référence, les tests d'hypothèses s'écrivent (pour $k \neq c$) :

$$H_0 : \theta_k = \theta_c$$

$$H_1 : \theta_k \neq \theta_c$$

Un test bilatéral ($H_1 : \theta_k > \theta_c$ ou $H_1 : \theta_k < \theta_c$) est également envisageable.

En nous basant toujours sur les sommes de rangs S_k , pour n suffisamment grand, les régions critiques s'écrivent

$$|S_k - S_c| \geq u_{1-a} \sqrt{\frac{nK(K+1)}{6}}$$

avec $a = \frac{\alpha}{2(K-1)}$, pour un test bilatéral.

Il devient

$$S_k - S_c \geq u_{1-a} \sqrt{\frac{nK(K+1)}{6}}$$

avec $a = \frac{\alpha}{K-1}$ pour le test unilatéral avec l'hypothèse alternative $H_1 : \theta_k > \theta_c$.

Exemple 42 (Les torches de soudure). Nous revenons sur les torches de soudage (page 146). Leur aspect avait été banalisé lors de l'expérimentation. En réalité, la torche n^o5 correspond à l'ancien modèle utilisé depuis toujours par les soudeurs. On cherche donc à savoir si les nouveaux modèles proposés (n^o1 à 4) s'en démarquent significativement. Ainsi, nous allons réitérer 4 tests dont l'hypothèse alternative est de la forme $H_1 : \theta_k > \theta_5$.

	SoudEUR	Traitement				
		Torche1	Torche2	Torche3	Torche4	Torche5
Bloc	1	3.9	4.1	4.2	4.1	3.3
	2	9.4	9.5	9.4	9	8.6
	3	9.7	9.3	9.3	9.2	8.4
	4	8.3	8	7.9	8.6	7.4
	5	9.8	8.9	9	9	8.3
	6	9.9	10	9.7	9.6	9.1

	Rangs				
	Torche1	Torche2	Torche3	Torche4	Torche5
1	2	3.5	5	3.5	1
2	3.5	5	3.5	2	1
3	5	3.5	3.5	2	1
4	4	3	2	5	1
5	5	2	3.5	3.5	1
6	4	5	3	2	1
S _k	23.5	22	20.5	18	6

n	6
K	5
alpha	0.01
a	0.0025
u _(1-a)	2.8070
Seuil	15.3748

Tableau des écarts		
Comparaisons	écart	Signif
1 vs. 5	17.5	oui
2 vs. 5	16	oui
3 vs. 5	14.5	non
4 vs. 5	12	non

Fig. 7.6. Comparaisons à une référence à l'issue du Test de Friedman - Torches de soudure

Les calculs sont résumés dans une feuille Excel (Figure 7.6) :

- À partir du tableau des données, nous produisons le tableau des rangs. Nous utilisons les rangs moyens pour les ex-aequo. Nous disposons de $n = 6$ blocs et $K = 5$ catégories.

- Nous calculons alors la somme des rangs par colonne S_k , nous obtenons $S_1 = 23.5$, $S_2 = 22$, $S_3 = 20.5$, $S_4 = 18$ et, la référence, $S_5 = 6$
- Le test de Friedman sur ces données était significatif à $\alpha = 1\%$. Puisque nous sommes dans un schéma unilatéral, nous déterminons

$$a = \frac{\alpha}{K-1} = \frac{0.01}{5-1} = 0.0025$$

- Le quantile de la loi normale d'ordre $1 - a$ associé est $u_{1-0.0025} = 2.8070$
- Le seuil critique à utiliser pour les comparaisons individuelles sera

$$seuil = u_{1-a} \sqrt{\frac{nK(K+1)}{6}} = 2.8070 \sqrt{\frac{6 \times 5 \times (5+1)}{6}} = 15.3748$$

- L'étape suivante consiste à former le "Tableau des écarts". La référence est la torche de soudeuse n^o5 , nous lui comparons les autres en formant l'écart ($e_k = S_k - S_5$). Nous obtenons ainsi $e_1 = 17.5$, $e_2 = 16$, $e_3 = 14.5$ et $e_4 = 12$.
- Les écarts significatifs, supérieurs au $seuil = 15.3748$, concernent les torches n^o1 et n^o2 . Les autres en revanche ne se démarquent pas significativement de la torche n^o5 .

Remarque 22 (Utiliser des valeurs critiques plus précises). Les tests que nous mettons en oeuvre pour tester les écarts de chaque groupe par rapport à la catégorie de référence ne sont pas indépendants. La quantité u_{1-a} utilisées dans les formules ci-dessus sont une approximation de valeurs critiques $q(\alpha, \#c)$ que l'on devrait mettre en oeuvre dans notre contexte. Les tables y afférentes sont disponibles dans quelques ouvrages de référence (voir par ex. [13], page *AIII*, page 321 ; $\#c$ correspond au nombre de comparaisons à réaliser, soit $\#c = K - 1$ dans notre cas). Si l'on se réfère à notre exemple ci-dessus, pour un test unilatéral, la véritable valeur critique serait $q(0.01, 3) = 2.77$, à comparer avec la valeur approchée $u_{1-a} = 2.8070$. L'écart reste raisonnable pour que l'on puisse se fier à l'approximation.

7.4 Tests pour les alternatives ordonnées

Exactement comme pour les tests pour échantillons indépendants (section 3.4), il est possible de préciser l'hypothèse alternative du test de Friedman en définissant un ordonnancement des paramètres de localisation. Le test d'hypothèses revient à opposer

$$H_0 : \theta_1 = \theta_2 = \dots \theta_K$$

$$H_1 : \theta_1 \leq \theta_2 \leq \dots \leq \theta_K$$

L'hypothèse nulle est rejetée lorsque une des inégalités de l'hypothèse alternative au moins est stricte.

Les exemples sont nombreux. Dans le cadre de mesures répétées, on peut citer le candidat au permis de conduire qui, à la suite de chaque séance de formation au code de la route, passe un questionnaire. On peut imaginer qu'à force d'apprentissage, il augmentera constamment la note obtenue. Dans le cadre des expérimentations en blocs aléatoires complets, on administre à l'intérieur d'un bloc des doses différentes d'un médicament pour évaluer son efficacité. En réalisant l'appariement, des individus semblables constitue un bloc, on diminue la variabilité due aux observations (voir la section 6.1).

7.4.1 Test de Page pour échantillons liés

Principe, statistique de test et région critique

Le tableau de données se présente comme celui analysé à l'aide du test de Friedman. Nous travaillons sur des données constituées de n blocs (lignes) et K traitements ordonnés (colonnes). L'ordre des colonnes du tableau ne peuvent pas être interverties, il doit être défini a priori sur la base de l'expérimentation qui a été menée (ex. dose croissante d'un médicament, prix croissant d'une catégorie de produit, etc.).

Pour réaliser le test, nous transformons les données en **rangs**, lesquels doivent être **définis à l'intérieur de chaque bloc**. Par rapport aux tests pour échantillons indépendants, cette spécificité nous permet d'exploiter au mieux l'appariement. Nous travaillons donc sur un tableau dont la structure a été décrite dans la section 7.2.1 (page 142). Nous exploiterons principalement les sommes de rangs par traitements S_k .

La statistique du test de Page pour échantillons liés s'écrit

$$L = \sum_{k=1}^K k \times S_k \quad (7.7)$$

De nouveau, nous avons une expression que l'on peut assimiler à une sorte de covariance entre le niveau de traitement k et la somme des rangs associés. Le test de Page pour échantillons appariés sert donc avant tout à caractériser l'évolution linéaire des paramètres de tendance centrale des distributions conditionnellement au niveau.

La région critique correspond aux grandes valeurs de L . Les seuils critiques sont lus dans des tables statistiques spécifiques, paramétrées par α , n et K (voir par exemple [13], page 354 et 255, table N).

Approximation normale pour les grands échantillons

Pour les grandes valeurs de n ou K , il est possible de passer par l'approximation normale. Les paramètres de la loi sous H_0 sont

$$E(L) = \frac{1}{4}nK(K+1)^2 \quad (7.8)$$

$$V(L) = \frac{nK^2(K^2-1)^2}{144(K-1)} \quad (7.9)$$

Le test est unilatéral, la région critique définie avec la statistique centrée et réduite

$$Z = \frac{L - E(L)}{\sqrt{V(L)}}$$

au risque α s'écrit

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

Exemple 43 (Pollution de la rivière). $K = 6$ stations de détection de la pollution (A, B, ..., F) ont été successivement placés le long d'une rivière qui traverse une ville. Des mesures ont été effectuées durant $n = 15$ jours. La variable d'intérêt est l'importance de la pollution. On cherche à savoir si au fil de son passage dans la ville, la pollution moyenne de l'eau augmente⁷.

Test de Page à partir de blocs aléatoires complets						
Bloc	Station					
	A	B	C	D	E	F
1	20	18	24	22	29	38
2	32	37	34	31	39	38
3	18	23	19	25	23	26
4	9	7	14	11	12	11
5	29	37	32	59	40	45
6	38	25	27	47	45	45
7	8	15	7	12	15	13
8	18	13	22	26	23	22
9	32	36	37	35	48	40
10	23	25	26	25	32	56
11	6	8	12	9	10	10
12	24	18	20	27	25	27
13	13	18	14	14	19	26
14	18	26	19	19	29	32
15	14	12	25	56	54	75

n	15
K	6

Rangs						
A	B	C	D	E	F	
2	1	4	3	5	6	
2	4	3	1	6	5	
1	3.5	2	5	3.5	6	
2	1	6	3.5	5	3.5	
1	3	2	6	4	5	
3	1	2	6	4.5	4.5	
2	5.5	1	3	5.5	4	
2	1	3.5	6	5	3.5	
1	3	4	2	6	5	
1	2.5	4	2.5	5	6	
1	2	6	3	4.5	4.5	
3	1	2	5.5	4	5.5	
1	4	2.5	2.5	5	6	
1	4	2.5	2.5	5	6	
2	1	3	5	4	6	
S_k	25	37.5	47.5	56.5	72	76.5

k	1	2	3	4	5	6
k x S_k	25	75	142.5	226	360	459

L	1287.5
E(L)	1102.50
V(L)	918.75
Z	6.1034
u 0.95	1.6449
p-value	5.2E-10

Fig. 7.7. Pollution de la rivière en ville - Test de Page pour échantillons liés

Nous travaillons à partir du tableau de données. Chaque ligne constitue un bloc (une journée d'observation dans notre cas), chaque colonne un niveau (une station, dans l'ordre de positionnement sur la rivière) (Figure 7.7) :

- Dans chaque ligne du tableau de données, nous calculons le rang des stations relativement à la valeur mesurée. Nous obtenons ainsi le tableau des rangs.
- Bien évidemment, par construction, la somme des rangs en ligne est égale à une constante $K(K + 1)/2 = 21$.
- Lorsqu'il y a des ex-aequo, nous utilisons la technique des rangs moyens qui nous est familière maintenant. Ainsi, pour la ligne $n^o 3$, nous avons les valeurs (après les avoir trié) $\{18, 19, 23, 23, 25, 26\}$, nous en déduisons les rangs $\{1, 2, 3.5, 3.5, 5, 6\}$. Nous reviendrons sur cet aspect un peu plus loin.
- Premier indicateur majeur dans cette analyse, la moyenne des rangs par colonne (par niveau) S_k . Si effectivement il y a augmentation de la pollution de l'eau durant son périple au milieu de la ville, la

⁷ Cet exemple est tiré de http://hdelboy.club.fr/Nonparam.htm#9bis_-_le_test_de_Page. Le problème y est décrit en italien, je me suis débrouillé avec le traducteur de Google, j'espère avoir traduit (suffisamment) fidèlement la teneur de l'expérimentation. L'intérêt pour nous est de pouvoir comparer nos calculs.

suite des S_k devrait être croissant. Ce qui est le cas ici. Il reste à valider cette première impression avec le calcul statistique.

- Pour cela, nous calculons la quantité intermédiaire $k \times S_k$, pour former la statistique

$$L = \sum_{k=1}^K k \times S_k = 25 + 75 + 142.5 + 226 + 360 + 459 = 1287.5$$

- n et K étant tous les deux assez grand, nous pouvons passer à l'approximation normale. Nous avons besoin alors de l'espérance mathématique et la variance de la statistique sous H_0 . L'espérance s'obtient avec

$$E(L) = \frac{1}{4}nK(K+1)^2 = \frac{1}{4}15 \times 6 \times (6+1)^2 = 1102.5$$

Et la variance

$$V(L) = \frac{nK^2(K^2-1)^2}{144(K-1)} = \frac{15 \times 6^2(6^2-1)^2}{144 \times (6-1)} = 918.75$$

- Nous en tirons la statistique centrée réduite

$$Z = \frac{L - E(L)}{\sqrt{V(L)}} = \frac{1287.5 - 1102.5}{\sqrt{918.75}} = 6.1034$$

- Au risque 5%, le seuil critique est $u_{0.95} = 1.6449$. Nous nous situons dans la région critique, nous rejetons l'hypothèse nulle. Effectivement, la pollution de l'eau augmente à mesure qu'elle avance dans la ville⁸.
- Un simple coup d'oeil sur le graphique croisant le niveau k (en abscisse) et les sommes de rangs conditionnellement au niveau S_k (en ordonnée) nous aurait permis de deviner la conclusion du test (Figure 7.8). On remarquera par ailleurs la forme linéaire de la tendance, caractéristique que retraduit parfaitement le test de Page.
- Il y avait des ex-aequo dans nos données, nous avons utilisé le principe des rangs moyens. En toute rigueur, *nous devrions également corriger la variance pour former une statistique Z ajustée*. Mais nous avons remarqué tout au long de ce support que : sauf cas très particuliers (très nombreux ex-aequo dans les blocs), la correction est minime, voire négligeable ; et elle va toujours dans le sens d'une plus grande significativité du test, notre conclusion ci-dessus ne saurait être remise en cause.

7.4.2 Test de Jonckheere pour échantillons liés

Principe, statistique de test et région critique

Curieusement, la catégorisation suivante est très présente dans la littérature : test pour alternatives ordonnées, échantillon indépendants, il s'agit du test de Jonckheere-Terpstra ; test pour alternatives ordonnées, échantillons liés, il s'agit du test de Page. Pourtant, tout comme il existe un test de Page pour

⁸ Nos résultats sont les mêmes que notre site source http://hdelboy.club.fr/Nonparam.htm#9bis_-_le_test_de_Page, c'est toujours rassurant. Mis à part une petite différence sur la p -value, mais avec un Z de 6.1034, la probabilité critique ne peut être que très faible, $p = 5.2 \times 10^{-10}$ pour notre part.

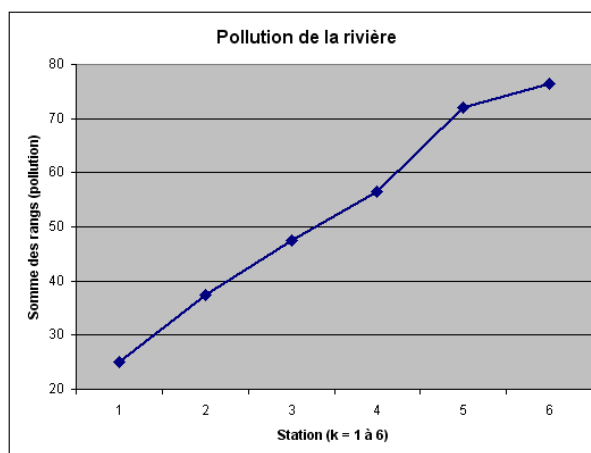


Fig. 7.8. Pollution de la rivière en ville - Relation entre k et les sommes de rang conditionnelles

échantillons indépendants (section 3.4.3), il existe également un test de Jonckheere pour échantillons liés, que nous présentons dans cette section (voir [3], pages 245 et 246).

Soit T_i la statistique de Mann et Whitney, définie sur un bloc

$$T_i = \sum_{k=1}^{K-1} \sum_{l=k+1}^K I(x_{ik} < x_{il}) \quad (7.10)$$

Pour résumer l'idée sous-jacente, nous dirons : pour chaque niveau k , nous comptabilisons le nombre d'observations plus grandes dans les niveaux qui lui sont supérieurs c.-à-d. $l = k + 1, \dots, K$. Il faut tenir compte des ex-aequo, pour y répondre formellement la fonction $I(\cdot)$ est définie comme suit :

$$I(x_{ik} < x_{il}) = \begin{cases} 1 & \text{si } x_{ik} < x_{il} \\ 0 & \text{si } x_{ik} > x_{il} \\ \frac{1}{2} & \text{si } x_{ik} = x_{il} \end{cases}$$

Si les valeurs sont effectivement ordonnées selon le niveau, T_i est au maximum, soit $K(K-1)/2$. Dans le cas contraire, l'ordonnancement est inversé par rapport au niveau, $T_i = 0$.

La statistique du test de Jonckheere pour échantillons liés est définie par

$$J = \sum_{i=1}^n T_i \quad (7.11)$$

La région critique correspond tout naturellement aux grandes valeurs de J .

Approximation normale

Lorsque n ou K sont assez grand, la distribution de la statistique du test est approximativement normale, de paramètres (sous H_0)

$$E(J) = \frac{1}{4}nK(K-1) \quad (7.12)$$

$$V(J) = \frac{1}{72}nK(K-1)(2K+5) \quad (7.13)$$

La statistique Z est formée selon notre habitude

$$Z = \frac{J - E(J)}{\sqrt{V(J)}}$$

Le test est unilatéral, la région critique au risque α est donc définie par

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

Exemple 44 (Pollution de la rivière revisitée). Reprenons l'exemple traité lors de la présentation du test de Page. Voyons ce qu'il en est avec le test de Jonckheere (Figure 7.9) :

Test de Jonckheere						
Bloc	Station					
	A	B	C	D	E	F
1	20	18	24	22	29	38
2	32	37	34	31	39	38
3	18	23	19	25	23	26
4	9	7	14	11	12	11
5	29	37	32	59	40	45
6	38	25	27	47	45	45
7	8	15	7	12	15	13
8	18	13	22	26	23	22
9	32	36	37	35	48	40
10	23	25	26	25	32	56
11	6	8	12	9	10	10
12	24	18	20	27	25	27
13	13	18	14	14	19	26
14	18	26	19	19	29	32
15	14	12	25	56	54	75

k	Comptage					
	1	2	3	4	5	6
4	4	2	2	1	0	
4	2	2	2	0	0	
5	2.5	3	1	1	0	
4	4	0	1.5	0	0	
5	3	3	0	1	0	
3	4	3	0	0.5	0	
4	0.5	3	2	0	0	
4	4	2.5	0	0	0	
5	3	2	2	0	0	
5	3.5	2	2	1	0	
5	4	0	2	0.5	0	
3	4	3	0.5	1	0	
5	2	2.5	2	1	0	
5	2	2.5	2	1	0	
4	4	3	1	1	0	

T
13
10
12.5
9.5
12
10.5
9.5
10.5
12
13.5
11.5
11.5
12.5
12.5
13

Somme	174
-------	-----

n	15
K	6

J	174
E(J)	112.50
V(J)	106.25

Z	5.9664
u_0.95	1.6449
p-value	1E-09

Fig. 7.9. Pollution de la rivière en ville - Test de Jonckheere pour échantillons liés

- A partir du tableau de données, nous construisons le tableau de comptage T_i .
- Voyons en détail la première ligne $i = 1$. Pour $k = 1$, nous avons $x_{11} = 20$, nous comptons 4 valeurs qui lui sont supérieures dans la ligne ($l = 2, \dots, K$), à savoir $\{24, 22, 29, 38\}$. Pour $k = 2$, $x_{12} = 18$, nous comptons 4 valeurs qui lui sont supérieures sur sa droite avec $\{24, 22, 29, 38\}$. Pour $k = 3$, $x_{13} = 24$, nous observons 2 valeurs supérieures parmi ($l = 3, \dots, K$), avec $\{29, 38\}$. Etc.

- En réalisant la somme de ces décomptes, nous obtenons $T_1 = 13$.
- Nous réalisons la même opération sur chaque ligne du tableau, nous obtenons au final la statistique de test en réalisant la somme des T_i

$$J = \sum_i T_i = 13 + 10 + 12.5 + 9.5 + \dots + 12.5 + 12.5 + 13 = 174$$

- Nous calculons alors l'espérance mathématique

$$E(J) = \frac{1}{4}nK(K-1) = \frac{1}{4}15 \times 6 \times (6-1) = 112.5$$

- Et la variance

$$V(J) = \frac{1}{72}nK(K-1)(2K+5) = \frac{1}{72}15 \times 6 \times (6-1) \times (2 \times 6 + 5) = 106.25$$

- Nous formons la statistique centrée et réduite

$$Z = \frac{J - E(J)}{\sqrt{V(J)}} = \frac{174 - 112.5}{\sqrt{106.25}} = 5.9664$$

- Pour un test à 5%, nous le confrontons avec le seuil critique $u_{0.95} = 1.6449$. Nous rejetons l'hypothèse nulle. Il y a bien un ordonnancement des paramètres de localisation selon k .
- On ne manquera pas de rapprocher ce résultat avec la valeur de Z pour le test de Page (Figure 7.7). Les valeurs obtenues pour les 2 tests sont très similaires. Il en est ainsi dans la très grande majorité des cas, même si le test de Page convient avant tout pour caractériser les tendances linéaires.
- Dernier commentaire, il y a des ex-aequo dans les lignes du tableau de données. On devrait introduire une correction de la variance. Mais elle s'avère généralement assez minime, et va dans le sens d'une augmentation de la significativité. Le résultat obtenu ici ne saurait être remis en cause.

Tests pour les variables binaires

Le cas des variables binaires $\{0, 1\}$ est un peu particulier. Nous pouvons les considérer comme des variables catégorielles (1 correspond à la présence d'un caractère), comme des variables ordinales (1 est mieux que 0), et à certains égards même, nous pouvons la considérer comme une variable quantitative (on peut calculer une moyenne, elle a un sens, elle correspond à la fréquence).

Dans ce chapitre nous nous positionnons dans le 1^{er} cas, X est une variable binaire catégorielle : $X = 1$ indique la présence de la caractéristique étudiée chez l'individu, $X = 0$ son absence. Nous travaillons sur des échantillons appariés ou des blocs aléatoires complets. La principale différence par rapport aux tests dédiés aux variables quantitatives ou ordinales (test des rangs signés de Wilcoxon, test de Friedman, etc.) est que l'amplitude de la différence $d_i = (x_{i1} - x_{i2})$ n'a pas de sens. De facto, le passage des données aux rangs est impossible. Nous utiliserons un autre canevas pour quantifier les disparités entre les populations. Il ne saurait être question de différences entre les paramètres de localisation ou d'échelle dans ce chapitre.

8.1 Test de McNemar pour la comparaison de $K = 2$ populations

8.1.1 Principe, statistique de test et région critique

Le test de McNemar s'applique très bien aux mesures de type "avant-après"¹. Dans une échantillon de taille n , on mesure l'apparition d'une caractéristique. Un événement survient. Chez les mêmes individus, on effectue de nouveau les mesures. **Il s'agit ni plus ni moins que d'un test de comparaison de proportions, sauf que les échantillons ne sont pas indépendants** puisque les mesures sont réalisées sur les mêmes individus. Le test paramétrique usuel de comparaison de proportions pour échantillons indépendants² n'est pas approprié dans ce cas.

Si π_1 (resp. π_2) est la probabilité $\Pr(X = 1)$ lors de la 1^{ère} (resp. 2^{nde}) mesure. Le test de McNemar peut s'écrire

¹ http://en.wikipedia.org/wiki/McNemar's_test

² Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf, chapitre 3

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 \neq \pi_2$$

Prenons un exemple simple, on a demandé à $n = 40$ des jeunes d'une ville s'ils sont "pour" (1) ou "contre" (0) les contrôles renforcés d'alcoolémie à la sortie des boîtes de nuit : 25 ont été pour, 15 contre. Après quelques semaines où effectivement les contrôles policiers ont été fortement présents, on leur a reposé la même question. Il s'avère que 32 sont maintenant "pour", et 8 "contre". On cherche à savoir si la proportion des "pour" a évolué. Manifestement certains individus qui étaient "contre" sont devenus "pour". Mais rien ne nous dit dans ces valeurs si le cas inverse s'est produit. Il se peut que certains jeunes, favorables aux contrôles initialement, se sont ravisés a posteriori pour différentes raisons. Il nous faut donc une structure plus riche pour exploiter les informations disponibles.

Il est d'usage de présenter les données dans un tableau de contingence du type

		Après		Total
		0	1	
Avant	0	a	b	a+b
	1	c	d	c+d
Total		a+c	b+d	n

A la marge, nous avons les informations sur les proportions de "1" avant les modifications $\hat{\pi}_1 = \frac{c+d}{n}$ et après $\hat{\pi}_2 = \frac{a+c}{n}$. Nous disposons des informations sur les flux de $0 \rightarrow 1$ (b) et de $1 \rightarrow 0$ (c). Les marges sont identiques si $(a + b = a + c)$ c.-à-d. $(b = c)$ pour les "0". De la même manière pour les "1", identité des marges $(c + d = b + d)$ entraîne $(c = d)$. Le test sera donc essentiellement fondé sur la confrontation entre les valeurs b et c .

Le test d'hypothèses peut s'écrire comme une comparaison de probabilités sous cet angle, on oppose

$$H_0 : \Pr(0 \rightarrow 1) = \Pr(1 \rightarrow 0)$$

$$H_1 : \Pr(0 \rightarrow 1) \neq \Pr(1 \rightarrow 0)$$

La statistique du test d'écrit alors

$$\chi^2 = \frac{(b - c)^2}{(b + c)} \quad (8.1)$$

Sous H_0 , elle est distribuée selon une loi du χ^2 à 1 degré de liberté. La région critique au risque α est tout naturellement

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

Remarque 23 (Correction de continuité). Sur les petits effectifs, surtout lorsque $(b + c)$ est faible puisque la statistique repose essentiellement dessus, certains conseillent d'introduire la correction de continuité de Yates (voir [13], page 76). Elle améliore l'approximation à l'aide de la loi du χ^2 . La statistique de test s'écrit alors

$$\chi^2 = \frac{(|b - c| - 1)^2}{(b + c)} \quad (8.2)$$

Nous avons aussi la possibilité, comme nous le verrons plus loin, de passer par les calculs exacts basés sur la loi binomiale.

Remarque 24 (D'autres applications). Le schéma "avant - après" est certainement le plus commode pour exposer le test de McNemar. Mais les applications sont plus larges. Il faut tout simplement que la variable d'intérêt soit binaire et que nous travaillions sur des échantillons appariés. Ainsi, le test s'applique également lorsque nous voulons confronter le jugement de 2 critiques gastronomiques qui doivent évaluer des plats préparés vendus dans des distributeurs automatiques. Ils doivent juger des séries de n plats en les étiquetant "bon pour la vente" ou non. Le test de McNemar permet de contrôler la cohérence de leur jugement.

Exemple 45 (Pour ou contre les contrôles d'alcoolémie). Calculons la statistique du test de McNemar à partir du tableau de contingence croisant les déclarations des jeunes avant et après les semaines de contrôles intensifs (Figure 8.1) :

		Après		Total
		contre	pour	
Avant	contre	3	12	15
	pour	5	20	25
	Total	8	32	40

KHI2	2.1176
ddl	1
KHI2_{0.95}	3.8415
p-value	0.14561

Fig. 8.1. Pour ou contre les contrôles d'alcoolémie à la sortie des boîtes de nuit - Test de McNemar

- Nous avons déjà cité les valeurs à la marge plus haut. Nous constatons maintenant que $b = 12$ individus qui étaient contre les contrôles, sont maintenant favorables. A l'inverse, $c = 5$ qui étaient pour sont maintenant opposés.
- Formons la statistique du χ^2 en introduisant la correction de continuité, nous obtenons

$$\chi^2 = \frac{(|b - c| - 1)^2}{(b + c)} = \frac{(|12 - 5| - 1)^2}{(12 + 5)} = 2.1176$$

- Au risque 5%, le seuil critique est défini par $\chi_{0.95}^2(1) = 3.8415$. Les données sont compatibles avec l'hypothèse nulle. Les réponses de nos jeunes ne sont pas significativement différentes aux deux dates étudiées.
- La probabilité critique du test est $p = 0.14561$

8.1.2 Un approche non symétrique du test de McNemar

Calcul exact de la probabilité critique

On remarque que la statistique du test repose uniquement sur les valeurs de b et c . Les quantités a et b n'interviennent pas, nous pouvons les considérer comme fixes.

Notons B la variable aléatoire associée à l'évènement ($0 \rightarrow 1$), que l'on mesure avec la quantité b dans le tableau de contingence. Les hypothèses à confronter dans le test de McNemar peuvent s'écrire de la manière suivante

$$\begin{aligned} H_0 : \Pr(B) &= \frac{1}{2} \\ H_1 : \Pr(B) &\neq \frac{1}{2} \end{aligned}$$

Sous H_0 , B est distribuée selon une loi binomiale de paramètres $\mathcal{B}(m, \frac{1}{2})$, où $m = b + c$. La probabilité critique du test peut être calculée à l'aide de l'expression

$$p = 2 \times \Pr_{\mathcal{B}(m, \frac{1}{2})} [B \geq \max(b, m - b)] \quad (8.3)$$

où

$$\begin{aligned} \Pr_{\mathcal{B}(m, \frac{1}{2})} [B \geq \max(b, m - b)] &= \sum_{l=\max(b, m-b)}^m \binom{m}{l} \left(\frac{1}{2}\right)^l \left(1 - \frac{1}{2}\right)^{(m-l)} \\ &= \sum_{l=\max(b, m-b)}^m \binom{m}{l} \left(\frac{1}{2}\right)^m \end{aligned}$$

Remarque 25 (Aile gauche et droite de la distribution de $\mathcal{B}(m, \frac{1}{2})$ pour un test bilatéral). La véritable probabilité critique pour un test bilatéral s'écrit

$$p = \Pr_{\mathcal{B}(m, \frac{1}{2})} [B \leq \min(b, m - b)] + \Pr_{\mathcal{B}(m, \frac{1}{2})} [B \geq \max(b, m - b)]$$

Mais parce que la loi $\mathcal{B}(m, \frac{1}{2})$ est symétrique, nous pouvons simplement calculer la probabilité sur une des ailes et la multiplier par 2, d'où l'expression 8.3.

Exemple 46 (Retour sur les contrôles d'alcoolémie). Dans notre exemple, $m = 17$ et $b = 12$. La probabilité critique basée sur la distribution binomiale s'écrit

$$\begin{aligned} p &= 2 \times \sum_{l=12}^{17} \binom{17}{l} \left(\frac{1}{2}\right)^{17} \\ &= 2 \times (0.04721 + 0.01816 + 0.00519 + 0.00104 + 0.00013 + 0.00001) \\ &= 2 \times 0.07173 \\ &= 0.14346 \end{aligned}$$

A rapprocher avec la probabilité critique égale à 0.14561 calculée à l'aide de l'approximation par la loi du χ^2 . Elle est de très bonne facture finalement.

Définir un test unilatéral

On se rend compte que la présentation usuelle du test de McNemar (section 8.1.1), un peu globalisée, masque les possibilités d'analyse fine des données. La formulation du test dans la section précédente nous éclaire mieux sur les informations que l'on peut extraire du tableau de contingence. Rien ne nous empêche en réalité de définir un test unilatéral du type

$$\begin{aligned} H_0 : \Pr(B) &= \frac{1}{2} \\ H_1 : \Pr(B) &\geq \frac{1}{2} \end{aligned}$$

La probabilité critique du test peut être alors calculée en se focalisant sur l'aile droite de la distribution de la loi binomiale, soit

$$p = \Pr_{\mathcal{B}(m, \frac{1}{2})} [B \geq b] \quad (8.4)$$

Remarque 26 (Approximation lorsque m est grand). Lorsque m est suffisamment élevé (en pratique³ $m > 10$), nous pouvons utiliser l'approximation normale. Nous formons la quantité Z , en introduisant la correction de continuité

$$Z = \frac{(b - \frac{m}{2}) - 0.5}{\sqrt{\frac{m}{4}}} = \frac{(2b - m) - 1.0}{\sqrt{m}} \quad (8.5)$$

La probabilité critique du test peut être obtenue avec

$$p' = 1 - \Phi(Z) = 1 - \Phi\left(\frac{(2b - m) - 1.0}{\sqrt{m}}\right) \quad (8.6)$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée et réduite.

³ Il est admis que l'approximation normale d'une loi binomiale $\mathcal{B}(m, \pi)$ est suffisamment bonne dès lors que $m \times \pi > 5$ et $m \times (1 - \pi) > 5$

Exemple 47 (Convaincre du bien-fondé des contrôles d'alcoolémie). On veut savoir si les contrôles effectués durant la période étudiée ont eu un impact positif sur les jeunes⁴. Nous sommes dans un cadre unilatéral où $H_1 : \Pr(B) \geq \frac{1}{2}$. La probabilité critique du test basé sur la loi binomiale, nous la connaissons déjà, nous l'avons calculée plus haut avec $\Pr(B \geq 12) = 0.07173$. Au risque 5%, les contrôles accrus n'ont pas eu d'impact positif sur les jeunes.

Voyons ce qu'il en est si nous utilisons l'approximation normale, Z est telle que

$$Z = \frac{(2 \times 12 - 17) - 1.0}{\sqrt{17}} = 1.45521$$

La probabilité critique obtenue avec la loi de répartition normale centrée réduite est

$$p' = 1 - \Phi(1.45521) = 1 - 0.92719 = 0.07281$$

Ici également, l'approximation normale est d'une très bonne tenue.

8.2 Le test de Stuart-Maxwell : généralisation de McNemar à L modalités

8.2.1 Principe, statistique de test et région critique

Le test de Stuart-Maxwell est une généralisation du test de McNemar pour les variables d'intérêt catégoriels à ($L \geq 2$) modalités (nominales ou ordinales). Lorsque $L = 2$ (variable binaire), les deux tests sont totalement équivalents (test de McNemar sans la correction de continuité).

Soient X_1 et X_2 les variables à L modalités mesurées. Nous travaillons toujours sur $K = 2$ échantillons appariés. Nous cherchons à savoir si la structure des réponses est différente. En termes probabilistes, si $\pi_{l.}$ (resp. $\pi_{.l}$) désigne la probabilité d'apparition de la modalité l pour la première variable (resp. seconde) mesurée, le test d'hypothèses s'écrit

$$H_0 : \pi_{1.} = \pi_{.1}, \pi_{2.} = \pi_{.2}, \dots, \pi_{l.} = \pi_{.l}, \dots, \pi_{L.} = \pi_{.L}$$

$$H_1 : \text{un des couples de probabilités au moins diffère}$$

Nous sommes dans un schéma de test d'équivalence distributionnelle, sauf que les échantillons ne sont pas indépendants, le test d'homogénéité⁵ du χ^2 n'est pas adapté.

Nous travaillons sur un tableau de contingence de taille $(L \times L)$, précisons les notations

⁴ On peut imaginer que c'est possible, si la maréchaussée est aimable et que personne n'a vomi sur les mains de l'affable fonctionnaire en se penchant par la portière pour souffler dans le ballon...

⁵ Voir http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf, chapitre 3.

.	X_2					.
X_1	1	...	l	...	L	Total
1	n_{11}	...	n_{1l}	...	n_{1L}	$n_{1.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
l	n_{l1}	...	n_{ll}	...	n_{lL}	$n_{l.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
L	n_{L1}	...	n_{Ll}	...	n_{LL}	$n_{L.}$
Total	$n_{.1}$...	$n_{.l}$...	$n_{.L}$	$n_{..} = n$

Si les réponses étaient totalement cohérentes, nous aurions $n_{ll} = n_{l.} = n_{.l}$ c.-à-d. toutes les valeurs seraient concentrées sur la diagonale principale. L'idée du test sera donc d'évaluer dans quelle mesure nous nous écartons de cette situation de référence.

Nous pouvons déduire de ces effectifs les fréquences absolues

$$f_{lt} = \frac{n_{lt}}{n}$$

Pour comparer les distributions marginales, nous formons le vecteur d'écart des fréquences marginales D de taille $(L-1) \times 1$ tel que

$$d_l = f_{l.} - f_{.l}$$

$L-1$ éléments suffisent car les L valeurs sont liées par une relation linéaire $\sum_l f_{l.} = \sum_l f_{.l} = 1$.

Soit S la matrice de variance covariance associée à d , de taille $(L-1) \times (L-1)$. Elle est composée des éléments suivants

$$\begin{aligned} s_{ll} &= f_{l.} + f_{.l} - 2 \times f_{ll} \\ s_{lt} &= -(f_{lt} + f_{tl}) \end{aligned}$$

Stuart et Maxwell⁶ proposent la statistique suivante pour confronter les distributions marginales

$$\chi_S^2 = nD'S^{-1}D \quad (8.7)$$

où D' est la transposée de D .

Sous H_0 , χ_S^2 est distribuée selon une loi du χ^2 à $(L-1)$ degrés de liberté. La région critique du test au risque α est donc définie par

$$R.C. : \chi_S^2 \geq \chi_{1-\alpha}^2(L-1)$$

⁶ Voir <http://ourworld.compuserve.com/homepages/jsuebersax/mcnemar.htm#stuart>

Exemple 48 (Vision oeil gauche - oeil droit). Cet exemple est repris d'un document décrivant la programmation de la méthode à l'aide d'une macro SAS⁷. Il semble avoir été tiré lui-même de l'article original de Stuart (1955). Il décrit la distance de vision chez les femmes âgées de 30 à 39 ans. On cherche à comparer les performances de l'oeil droit et de l'oeil gauche. Il n'est pas question de données de type "avant-après" ici. Mais indéniablement, nous sommes dans un schéma de comparaison de distributions sur échantillons appariés. C'est le même concept qui est mesuré en ligne et en colonne du tableau de contingence. La variable d'intérêt est ordinaire dans cet exemple, mais la technique s'applique tout aussi bien aux variables nominales.

Tableau de données

		Left Eye				
		Highest	Second	Third	Lowest	Total
Right Eye	Highest	1520	266	124	66	1976
	Second	234	1512	432	78	2256
	Third	117	362	1772	205	2456
	Lowest	36	82	179	492	789
	Total	1907	2222	2507	841	7477

Tableau de fréquences

		Left Eye				
		Highest	Second	Third	Lowest	Total
Right Eye	Highest	0.2033	0.0356	0.0166	0.0088	0.2643
	Second	0.0313	0.2022	0.0578	0.0104	0.3017
	Third	0.0156	0.0484	0.2370	0.0274	0.3285
	Lowest	0.0048	0.0110	0.0239	0.0658	0.1055
	Total	0.2550	0.2972	0.3353	0.1125	1.0000

Test de Stuart-Maxwell

D

0.0092

0.0045

-0.0068

S

0.1127

-0.0669

-0.0322

-0.0669

0.1945

-0.1062

-0.0322

-0.1062

0.1898

S⁻¹

18.5574

11.6678

9.6805

11.6678

14.7411

10.2300

9.6805

10.2300

12.6375

KHI2

11.9566

ddl

3

KHI2 (0.95)

7.8147

p-value

0.00753

- Nous inversons la matrice de variance covariance pour obtenir S^{-1}
- Il ne nous reste plus qu'à construire la forme quadratique, soit

$$\chi_S^2 = nD'S^{-1}D = 7477 \times D'S^{-1}D = 11.9566$$

- Le degré de liberté est $L - 1 = 4 - 1$, le seuil critique du test au risque $\alpha = 5\%$ est $\chi_{0.95}^2(3) = 7.8147$. Nous sommes dans la région critique. La qualité de vision de l'oeil gauche et de l'oeil droite ne sont pas du même ordre.
- La probabilité critique du test est $p = 0.00753$

8.2.2 La variante de Bhapkar

Une alternative, généralement plus puissante, du test de Stuart-Maxwell a été développée. Il s'agit du test de Bhapkar (1966). La procédure est exactement la même, à l'exception de l'estimation de la matrice de variance covariance de D . Nous la noterons B dans cette section, ses valeurs sont obtenues à l'aide des relations suivantes⁸ :

$$\begin{aligned} b_{ll} &= f_{l.} + f_{.l} - 2 \times f_{ll} - (f_{.l} - f_{l.})^2 \\ b_{lt} &= -(f_{lt} + f_{tl}) - (f_{.l} - f_{l.})(f_{.t} - f_{t.}) \end{aligned}$$

La statistique du test s'écrit

$$\chi_B^2 = nD'B^{-1}D \quad (8.8)$$

Sous H_0 , elle est distribuée également selon une loi du χ^2 à $(L - 1)$ degrés de liberté.

Remarque 27 (Déduire la statistique de Bhapkar à partir de celui de Stuart-Maxwell). Il est possible de trouver un passage entre les 2 statistiques, elle s'écrit simplement

$$\chi_B^2 = \frac{\chi_S^2}{1 - \frac{\chi_S^2}{n}} \quad (8.9)$$

Exemple 49 (L'oeil du talion). Reprenons notre exemple "oeil gauche - oeil droit" ci dessus (Figure 8.2) et appliquons lui le test de Bhapkar en utilisant la démarche standard. Par la suite nous vérifierons si la formule de passage est effectivement opérante.

La feuille de calcul vient à la suite de la précédente (Figure 8.3) :

- Les tableaux de données et de fréquences sont les mêmes. Le vecteur D n'est donc pas modifié.

⁸ Voir <http://www2.sas.com/proceedings/forum2008/382-2008.pdf>, équations 7 et 8.

Test de Bhapkar																								
D	<table><tr><td>0.0092</td></tr><tr><td>0.0045</td></tr><tr><td>-0.0068</td></tr></table>				0.0092	0.0045	-0.0068																	
0.0092																								
0.0045																								
-0.0068																								
B	<table><tr><td>0.1127</td><td>-0.0669</td><td>-0.0322</td></tr><tr><td>-0.0669</td><td>0.1944</td><td>-0.1062</td></tr><tr><td>-0.0322</td><td>-0.1062</td><td>0.1897</td></tr></table>			0.1127	-0.0669	-0.0322	-0.0669	0.1944	-0.1062	-0.0322	-0.1062	0.1897	B ⁻¹	<table><tr><td>18.5825</td><td>11.6844</td><td>9.6883</td></tr><tr><td>11.6844</td><td>14.7521</td><td>10.2352</td></tr><tr><td>9.6883</td><td>10.2352</td><td>12.6400</td></tr></table>		18.5825	11.6844	9.6883	11.6844	14.7521	10.2352	9.6883	10.2352	12.6400
0.1127	-0.0669	-0.0322																						
-0.0669	0.1944	-0.1062																						
-0.0322	-0.1062	0.1897																						
18.5825	11.6844	9.6883																						
11.6844	14.7521	10.2352																						
9.6883	10.2352	12.6400																						
KHI2	11.9757																							
ddl	3																							
KHI2 (0.95)	7.8147																							
	p-value		0.00747																					
Vérification formule de passage																								
Stuart-Maxwell	11.9566																							
Bhapkar	11.9757																							

Fig. 8.3. Comparaison vision oeil gauche - oeil droit - Test de Bhapkar

- Il nous faut calculer la matrice de variance covariance B . Prenons l'exemple de 2 cellules

$$b_{11} = f_{1.} + f_{.1} - 2 \times f_{11} - (f_{.1} - f_{1.})^2 = 0.2643 + 0.2550 - 2 \times 0.2033 - (0.2550 - 0.2643)^2 = 0.1127$$

$$b_{12} = -(f_{12} + f_{21}) - (f_{.1} - f_{1.})(f_{.2} - f_{2.}) = -(0.0356 + 0.0313) - (0.2550 - 0.2643)(0.2972 - 0.3017) = -0.0669$$

- Nous inversons la matrice B , nous formons la statistique de test

$$\chi_B^2 = nD'B^{-1}D = 7477 \times D'B^{-1}D = 11.9757$$

- Au risque 5%, nous devons la comparer à $\chi_{0.95}^2(3) = 7.8147$. Nous sommes dans la région critique, la conclusion du test est cohérente avec le test de Stuart-Maxwell. La probabilité critique du test ici est $p = 0.00747$.
- On notera que les 2 approches proposent une estimation très proche de la matrice de variance covariance sur notre exemple. Il est naturel par la suite que la statistique de test soit également très similaire.
- Voyons maintenant ce qu'il en est au niveau de la formule de passage (voir "Vérification formule de passage" dans la feuille EXCEL), si nous calculons la quantité

$$\chi_B^2 = \frac{\chi_S^2}{1 - \frac{\chi_S^2}{n}} = \frac{11.9566}{1 - \frac{11.9566}{7477}} = 11.9757$$

La valeur obtenue coïncide exactement avec celle produite à l'aide de l'approche directe ci-dessus. On peut effectivement se servir aisément des résultats du test de Stuart-Maxwell pour déduire ceux de Bhapkar.

8.3 Test Q de Cochran pour $K \geq 2$ populations

8.3.1 Principe, statistique de test et région critique

Le test Q de Cochran est une généralisation du test de McNemar où l'on traite $K \geq 2$ échantillons appariés, dans un plan d'expérience en bloc aléatoire complet. Il s'applique aussi au cas des mesures répétées c.-à-d. un seul échantillon dans laquelle une variable a été mesurée K fois. La variable d'intérêt est binaire. On peut le voir comme une variante du test de Friedman.

Nous disposons de K mesures binaires, prenant leurs valeurs dans $\{1, 0\}$. Nous les noterons X_k , $k = 1, \dots, K$. Il s'agit de réaliser un test de comparaison de proportions

$$H_0 : \pi_1 = \dots = \pi_K$$

$$H_1 : \text{une au moins diffère des autres}$$

où π_k représente la probabilité d'occurrence de la valeur "1" pour la variable X_k : $\pi_k = \Pr(X_k = 1)$

Nous travaillons sur un tableau de données de la forme suivante

.	Traitement					.
Bloc	X_1	\dots	X_k	\dots	X_K	Somme
1	x_{11}	\dots	x_{1k}	\dots	x_{1K}	L_1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	x_{i1}	\dots	x_{ik}	\dots	x_{iK}	L_i
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	x_{n1}	\dots	x_{nk}	\dots	x_{nK}	L_n
Somme	C_1	\dots	C_k	\dots	C_K	S

où

- x_{ik} prend la valeur 1 ou 0 ;
- $C_k = \sum_i x_{ik}$ est le nombre d'apparition de la valeur 1 dans la k^{eme} colonne ;
- $L_i = \sum_k x_{ik}$ est le nombre d'apparition de la valeur 1 dans la i^{eme} ligne ;
- $S = \sum_{k=1}^K C_k = \sum_{i=1}^n L_i = \sum_i \sum_k x_{ik}$ est la somme totale des valeurs du tableau.

La statistique du test de Cochran est définie de la manière suivante⁹

$$Q = K(K-1) \frac{\sum_{k=1}^K (C_k - \frac{S}{K})^2}{\sum_{i=1}^n L_i (K - L_i)} \quad (8.10)$$

Cette formule nous éclaire sur le mode de fonctionnement du test. Si $\pi_k = \pi$, $\forall k$ c.-à-d. tous les traitements se valent, nous aurons, aux fluctuations d'échantillonnage près, $C_k = \frac{S}{K}$. Le

⁹ Voir http://en.wikipedia.org/wiki/Cochran_test ; voir aussi <http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/cochran.htm>

numérateur de l'équation 8.10 sera égal à 0, il en sera de même pour Q . Plus les traitements induiront des effets différents, plus la disparité des C_k autour de $\frac{S}{K}$ sera grande, induisant une augmentation de Q .

Pour être complet, il faut introduire la disparité des valeurs dans les lignes du tableau. Certains individus (blocs) ont plus d'aptitude que d'autres à avoir le caractère "1". A l'extrême, un individu a toujours la valeur "1" quel que soit le traitement mis en oeuvre, de fait $L_i = K$; d'autres en revanche n'auront jamais le caractère "1", dans ce cas $L_i = 0$. Il faut en tenir compte. Ce qui est fait au dénominateur de la statistique Q de Cochran.

On retrouve également l'écriture suivante dans la littérature (voir [13], page 171)

$$Q = \frac{(K-1) \left[K \sum_{k=1}^K C_k^2 - \left(\sum_{k=1}^K C_k \right)^2 \right]}{K \sum_{i=1}^n L_i - \sum_{i=1}^n L_i^2} \quad (8.11)$$

Sous H_0 , Q suit asymptotiquement une loi du χ^2 à $(K-1)$ degrés de liberté. L'approximation est de bonne qualité dès que pour $(n \geq 4)$ et $(n \times K > 24)$. La région critique du test au risque α s'écrit naturellement

$$R.C. : Q \geq \chi_{1-\alpha}^2(K-1)$$

Exemple 50 (Difficultés pour alpinistes). $K = 3$ ascensions différentes sont tentées par $n = 5$ alpinistes. Les succès sont notés "1", les échecs "0". On cherche à savoir si les ascensions sont de difficultés égales. En d'autres termes, on veut comparer la probabilité de réussir les différents types d'ascensions¹⁰.

Individu	Ascension 1	Ascension 2	Ascension 3	L_i	L_i^2	
1	1	1	0	2	4	
2	1	0	1	2	4	
3	0	0	1	1	1	
4	0	1	1	2	4	
5	1	0	1	2	4	
C_k	3	2	4	9	17	Somme
C_k^2	9	4	16			

n	5
K	3

Numérateur	12
Dénominateur	10

Q	1.2
---	-----

ddl	2
-----	---

KHI2 (0.95)	5.9915
-------------	--------

p-value	0.5488
---------	--------

Fig. 8.4. Comparer les difficultés de $K = 3$ ascensions - Test Q de Cochran

Les données et les calculs sont résumés dans une feuille Excel (Figure 8.4) :

¹⁰ Ce exemple est extrait du site <http://www.cons-dev.org/elearning/ando/04/44/44.html>

- A partir du tableau de données, nous calculons les marges en ligne L_i (ex. $L_1 = 1 + 1 + 0 = 2$) et leur carré L_i^2 (ex. $L_1^2 = 2^2 = 4$).
- Nous pouvons calculer les sommes $\sum_i L_i = 9$ et $\sum_i L_i^2 = 17$.
- De même, pour les colonnes, nous obtenons C_i (ex. $C_1 = 1 + 1 + 0 + 0 + 1 = 3$) et C_i^2 (ex. $C_1^2 = 3^2 = 9$).
- Nous pouvons maintenant former la statistique du test à l'aide de l'équation 8.11

$$\begin{aligned}
 Q &= \frac{(K-1) \left[K \sum_{k=1}^K C_k^2 - \left(\sum_{k=1}^K C_k \right)^2 \right]}{K \sum_{i=1}^n L_i - \sum_{i=1}^n L_i^2} \\
 &= \frac{(3-1) \left[3(9+4+16) - (3+2+4)^2 \right]}{3 \times 9 - 17} = \frac{2 \times [87 - 81]}{10} = \frac{12}{10} \\
 &= 1.2
 \end{aligned}$$

- Sous H_0 , Q suit une loi du χ^2 à $(K-1 = 3-1 = 2)$ degrés de liberté. Le seuil critique du test au risque $\alpha = 5\%$ est $\chi_{0.95}^2(2) = 5.9915$. Les données sont compatibles avec l'hypothèse nulle. Les ascensions sont de difficultés égales.
- La probabilité critique du test est $p = 0.5488$

8.3.2 Sorties des logiciels

Bloc	Interview1	Interview2	Interview3
1	0	0	0
2	1	1	0
3	0	1	0
4	0	0	0
5	1	0	0
6	1	1	0
7	1	1	0
8	0	1	0
9	1	0	0
10	0	0	0
11	1	1	1
12	1	1	1
13	1	1	0
14	1	1	0
15	1	1	0
16	1	1	1
17	1	1	0
18	1	1	0

Fig. 8.5. Influence du mode d'interview sur les réponses - Test Q de Cochran

Pour décrire les sorties des logiciels, nous reprenons un exemple de Siegel et Castellan ([13], page 173; figure 8.5). On a interrogé $K = 3$ échantillons de $n = 18$ familles, elles devaient répondre "oui" (1) ou "non" (2) à une question. Dans chaque échantillon, l'interview a été menée d'une manière différente (enthousiaste, courtoise, détachée). On veut vérifier que la proportion de "oui" est la même dans chaque échantillon c.-à-d. on veut savoir si la manière de mener une interview influence les réponses. Une caractéristique forte des données est qu'elles ont été organisées en $n = 18$ blocs de 3 familles. Les familles

situées dans le même bloc sont similaires au regard de l'étude. Nous sommes exactement dans un plan d'expérience en blocs aléatoires complets. Le test de Cochran doit être préféré à une comparaison de proportions pour échantillons indépendants.

Results

Variables		Cochran's Q Statistic	
Att.	Sum(Values)	Stat.	Value
Interview1	13	Cochran's Q	16.66667
Interview2	13	d.f.	2
Interview3	3	p-value	0.00024

Fig. 8.6. Influence du mode d'interview sur les réponses - Test Q de Cochran avec TANAGRA

Le logiciel **TANAGRA** fournit les résultats suivants (Figure 8.6). Nous y retrouvons :

- La somme des valeurs par variable c.-à-d. le nombre de réponses positives puisque les variables sont binaires 0/1. Nous avons ainsi $C_1 = 13$, $C_2 = 13$ et $C_3 = 3$
- La statistique de Cochran est produite, $Q = 16.66667$.
- Sous h_0 , elle suit une loi du χ^2 à $(K - 1 = 3 - 1 = 2)$ degrés de liberté. La probabilité critique du test est $p = 0.00024$. Pas de doute, la manière de poser les questions pèse sur les réponses.

Bien entendu, le test Q de Cochran est disponible dans d'autres logiciels. Pour SPSS par exemple, on se rapportera volontiers au descriptif proposé par Garson sur son site (voir <http://www2.chass.ncsu.edu/garson/PA765/friedman.htm> - Cochran's Q Test).

8.3.3 Détecter la source des écarts

Si le test de Cochran aboutit au rejet de l'hypothèse nulle, nous savons que deux des proportions au moins sont différentes. L'étape suivante de l'étude est souvent la détection de ces écarts en testant toutes les configurations possibles. Nous sommes dans un schéma de comparaisons multiples. Nous mettons en place $\frac{K(K-1)}{2}$ tests bilatéraux individuels : pour $k \neq l$,

$$H_0 : \pi_k = \pi_l$$

$$H_1 : \pi_k \neq \pi_l$$

La région critique est s'écrit :

$$|\hat{\pi}_k - \hat{\pi}_l| \geq u_{1-\alpha} \sqrt{2 \times \left[\frac{K \sum_i L_i - \sum_i L_i^2}{n^2 K(K-1)} \right]} \quad (8.12)$$

avec

$$a = \frac{\alpha/2}{K(K-1)/2} = \frac{\alpha}{K(K-1)}$$

Est définie de manière à conserver le niveau du risque global α du test Q de Cochran.

Remarque 28 (Autre stratégie de recherche des écarts). Parfois, nous souhaitons uniquement nous intéresser à certains écarts, définis *a priori*. Nous pouvons aussi mettre en place des tests unilatéraux. Le schéma reste le même, la définition du risque individuel a , et donc de la quantité u_{1-a} , est modifiée en revanche (voir [12], pages 870 à 872).

Exemple 51 (Choix de véhicules). On a demandé à $n = 12$ femmes si elles étaient d'accord pour l'achat de véhicule d'une marque spécifique. $K = 3$ marques ont été proposées [Chenescio (A), Howasaki (B) et Gemini(C)], à chaque fois la personne interrogée devait répondre oui (1) ou non (0) (voir [12], page 868).

Nous sommes typiquement dans un schéma de mesures répétées, avec une réponse binaire. Pour savoir si les personnes sont plus attirées par une des marques en particulier, le test Q de Cochran est tout à fait indiqué. Les calculs sont retracés dans une feuille EXCEL (Figure 8.7, I et II) :

Tableau de données et calculs intermédiaires (I)					
i	Chenescio (A)	Howasaki (B)	Gemini (C)	L_i	L²_i
1	1	1	0	2	4
2	0	1	0	1	1
3	1	1	1	3	9
4	0	1	0	1	1
5	0	1	0	1	1
6	0	1	1	2	4
7	0	0	0	0	0
8	0	1	0	1	1
9	1	1	0	2	4
10	0	1	0	1	1
11	0	0	0	0	0
12	0	0	1	1	1
C_k	3	9	3	15	27
C²_k	9	81	9	Somme	

pi^k	0.25	0.75	0.25
------	------	------	------

Test Q de Cochran (II)	
n	12
K	3
Numérateur	144
Dénominateur	18
Q	8
alpha	0.05
Q 0.95(2)	5.99

Source des écarts (III)	
a	0.0083
u	2.3940
B	0.2041
seuil	0.4887

Tableau des écarts		
Ecart	pi_k - pi_i	Significatif
A vs. B	0.5	oui
A vs. C	0	non
B vs. C	0.5	oui

Fig. 8.7. Préférences de marques de véhicules - Test Q de Cochran et comparaisons multiples

– Nous calculons les marges en lignes L_i et en colonnes C_k . Nous les passons au carré.

- Dans chaque colonne, nous pouvons former la proportion estimée $\hat{\pi}_k$ que nous opposerons dans l'étape suivante. Ici, $\hat{\pi}_1 = 0.25$, $\hat{\pi}_2 = 0.75$ et $\hat{\pi}_3 = 0.25$.
- Nous formons la statique de Cochran

$$\begin{aligned}
 Q &= \frac{(K-1) \left[K \sum_{k=1}^K C_k^2 - \left(\sum_{k=1}^K C_k \right)^2 \right]}{K \sum_{i=1}^n L_i - \sum_{i=1}^n L_i^2} \\
 &= \frac{(3-1) \left[3(9+81+9) - (3+9+3)^2 \right]}{3 \times 15 - 27} = \frac{144}{18} \\
 &= 8
 \end{aligned}$$

- Pour risque $\alpha = 5\%$, le seuil critique du test est $\chi_{0.95}^2(2) = 5.99$. Puisque $Q = 8 > 5.99$, nous rejetons l'hypothèse d'égalité des proportions des réponses positives.

Nous aboutissons au rejet de l'hypothèse nulle, il est entendu qu'une des proportions au moins est différente des autres (ou d'un autre). Maintenant, essayons de détecter justement les écarts significatifs en testant toutes les configurations possibles, nous mettons en place les tests individuels (Equation 8.12), nous obtenons les résultats suivants (Figure 8.7, III) :

- Tout d'abord, nous calculons le risque individuel $a = \frac{\alpha}{K(K-1)} = \frac{0.05}{3(3-1)} = 0.0083$.
- Nous en déduisons $u_{1-0.0083} = 2.3940$.
- Formons la quantité intermédiaire B , avec

$$\begin{aligned}
 B &= \sqrt{2 \times \left[\frac{K \sum_i L_i - \sum_i L_i^2}{n^2 K(K-1)} \right]} \\
 &= \sqrt{2 \times \left[\frac{3 \times 15 - 27}{12^2 \times 3 \times (3-1)} \right]} \\
 &= 0.2041
 \end{aligned}$$

Le seuil critique pour les tests individuels est égal à

$$u_{1-0.0083} \times B = 2.3940 \times 0.2041 = 0.4887$$

- Tout écart (en valeur absolue) supérieur au seuil 0.4887 sera considéré comme significatif, c'est le cas pour les oppositions "A vs. B" ($|\hat{\pi}_1 - \hat{\pi}_2| = 0.5$) et "B vs. C" ($|\hat{\pi}_2 - \hat{\pi}_3| = 0.5$).

Gestion des versions

La première version (version 1.0) de ce fascicule a été finalisée fin août 2008. Il comprend 8 chapitres :

1. Tests génériques de comparaison de $K = 2$ échantillons.
2. Tests de rang dans un modèle de localisation pour $K = 2$ échantillons.
3. Tests de rang dans un modèle de localisation pour $K \geq 2$ échantillons.
4. Tests de rang dans un modèle d'échelle.
5. Retour sur les statistiques de rang linéaires.
6. Tests pour $K = 2$ échantillons liés.
7. Tests pour $K \geq 2$ échantillons liés.
8. Tests pour les variables binaires.

B

Les tests non paramétriques avec le logiciel TANAGRA

TANAGRA (<http://eric.univ-lyon2.fr/~ricco/tanagra/>) est un logiciel de data mining, statistique et analyse de données *open source*, totalement gratuit. La première version a été mise en ligne en Janvier 2004. Plusieurs techniques non paramétriques ont été implémentées à l'été 2005 (voir la section "Nouveautés" de notre site web http://eric.univ-lyon2.fr/~ricco/tanagra/fr/contenu_nouveautes.html).

La dernière version de TANAGRA (version 1.4.27, été 2008) accompagne ce document. Les sorties des techniques existantes ont été très légèrement remaniées. De nouvelles techniques ont été introduites. Les procédures sont regroupées dans l'onglet "NONPARAMETRIC STATISTICS" du logiciel (Figure B.1).

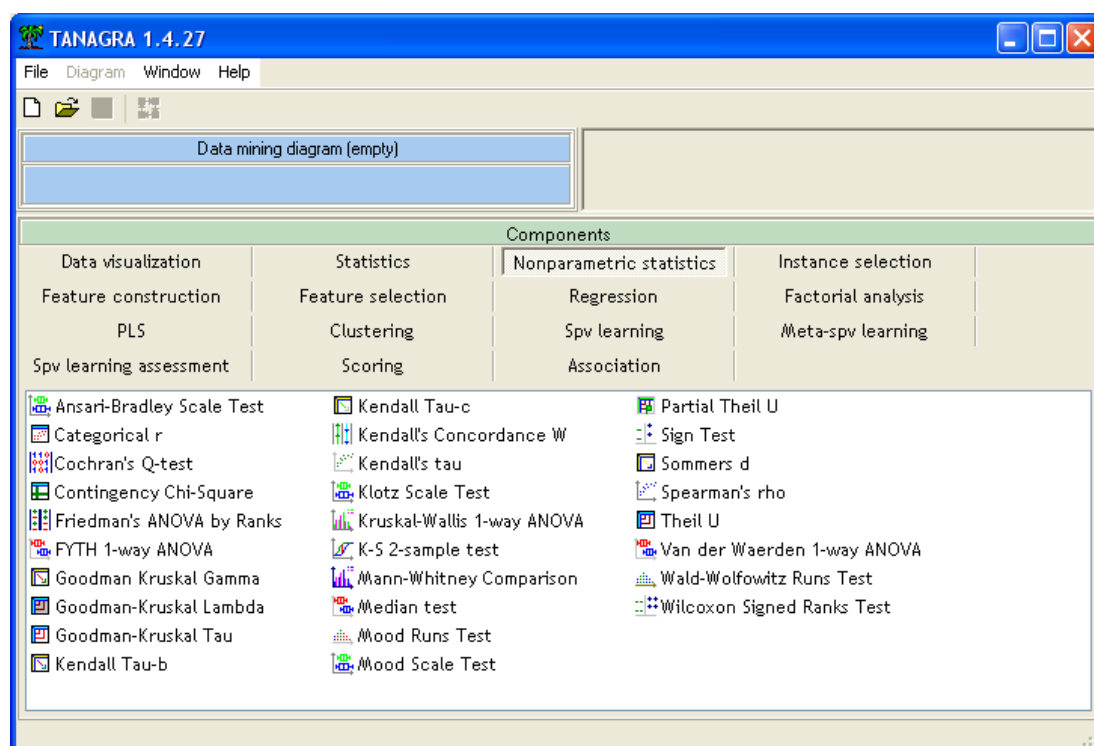


Fig. B.1. Liste des tests non paramétriques dans TANAGRA - Version 1.4.27

Concernant le détail de la mise en oeuvre des techniques à l'aide du logiciel, notamment la préparation et l'accès aux données, plusieurs **tutoriels** sont disponibles. Ils sont regroupés sur un site dédié <http://tutoriels-data-mining.blogspot.com/>.

S'agissant des tutoriels en relation directe avec ce fascicule, nous citerons volontiers :

Tests non paramétriques de comparaison de population, il décrit la mise en oeuvre des principaux tests présentés dans ce document

<http://tutoriels-data-mining.blogspot.com/2008/04/tests-non-parametriques.html>

Analyse de variance de Friedman

<http://tutoriels-data-mining.blogspot.com/2008/04/analyse-de-variance-de-friedman.html>

La mise en oeuvre des *tests paramétriques* est décrite dans plusieurs tutoriels, plus particulièrement <http://tutoriels-data-mining.blogspot.com/2008/07/comparaison-de-populations-tests.html>

D'autres viendront au fur et à mesure...

Les tests non paramétriques avec d'autres logiciels libres

Les tests non paramétriques sont bien entendu implémentés dans d'autres logiciels libres. On pense avant tout à **R** (<http://www.r-project.org/>), que nous utilisons d'ailleurs pour illustrer certaines méthodes. Mais d'autres logiciels proposent également ces techniques. Parmi les logiciels disponibles, **OpenStat** de Bill Miller a retenu mon attention (<http://www.statpages.org/miller/openstat/>).

C'est un travail remarquable. Le programme compilé peut être facilement installé. La prise en main est assez rapide, OpenStat est piloté par menu. L'importation des données ne pose pas de problèmes particulier (format texte, séparateur tabulation). Un guide de l'utilisateur et un manuel de référence accompagnent le logiciel. Enfin, cerise sur le gâteau, le code source est en ligne. Il est toujours instructif de voir comment les autres s'y prennent pour organiser leurs procédures. A titre indicatif, nous restituons une copie d'écran montrant le contenu du sous-menu "Nonparametric" du logiciel (Figure C.1). Consulter régulièrement le site permet de suivre les derniers développements.

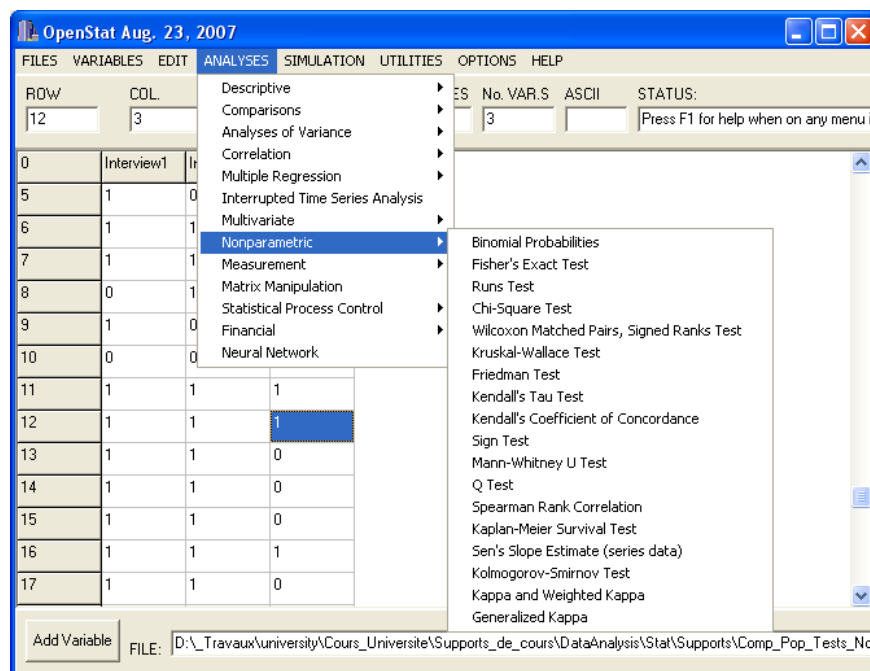


Fig. C.1. Liste des tests non paramétriques dans OpenStat

D

Tables statistiques

De nombreuses tables statistiques éparées sont disponibles sur internet. Fort heureusement, certains sites font l'effort de les réunir. Nous conseillerons en priorité les références suivantes concernant les tests non paramétriques :

- <http://courses.wcupa.edu/rbove/eco252/252suppkey.htm>
- <http://www.york.ac.uk/depts/maths/tables/pdf.htm>
- <http://www.cons-dev.org/elearning/ando/tables/tablesH.html>

Quelques tables sont reprises dans ces annexes. Elles accompagnent les techniques que nous présentons.

D.1 Test de Kolmogorov-Smirnov

Table de Kolmogorov-Smirnov pour le test d'homogénéité de 2 échantillons : valeurs critiques de D_{n_1, n_2} pour $n_1, n_2 \leq 20$ avec $n_1 \neq n_2$ (Figure D.1).

$\begin{smallmatrix} n_1 \\ n_2 \end{smallmatrix}$	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5		0.8	0.8	0.75	0.778	0.8	0.709	0.717	0.692	0.657	0.733	0.8	0.647	0.667	0.642	0.65
6	1		0.714	0.708	0.722	0.667	0.652	0.667	0.667	0.643	0.533	0.625	0.667	0.667	0.614	0.6
7	1	0.857		0.714	0.667	0.657	0.623	0.631	0.615	0.643	0.59	0.571	0.571	0.571	0.571	0.564
8	0.875	0.833	0.857		0.639	0.6	0.602	0.625	0.596	0.571	0.558	0.625	0.566	0.611	0.539	0.55
9	0.889	0.883	0.778	0.764		0.589	0.596	0.583	0.556	0.556	0.556	0.542	0.536	0.556	0.52	0.517
10	0.9	0.8	0.757	0.75	0.7		0.545	0.55	0.569	0.529	0.533	0.525	0.524	0.511	0.495	0.55
11	0.818	0.818	0.766	0.727	0.707	0.7		0.545	0.524	0.532	0.509	0.506	0.497	0.49	0.488	0.486
12	0.833	0.833	0.714	0.708	0.694	0.667	0.652		0.519	0.512	0.517	0.5	0.49	0.5	0.474	0.483
13	0.8	0.769	0.714	0.692	0.667	0.646	0.636	0.608		0.489	0.492	0.486	0.475	0.47	0.462	0.462
14	0.8	0.762	0.786	0.679	0.667	0.643	0.623	0.619	0.571		0.467	0.473	0.467	0.46	0.455	0.45
15	0.8	0.767	0.714	0.675	0.667	0.667	0.618	0.6	0.59	0.586		0.475	0.455	0.456	0.446	0.45
16	0.8	0.75	0.688	0.688	0.653	0.625	0.602	0.604	0.582	0.563	0.554		0.456	0.444	0.437	0.437
17	0.8	0.716	0.706	0.647	0.647	0.624	0.588	0.583	0.576	0.563	0.557	0.526		0.435	0.437	0.429
18	0.788	0.778	0.69	0.653	0.667	0.6	0.596	0.583	0.585	0.556	0.544	0.535	0.536		0.415	0.422
19	0.747	0.728	0.684	0.645	0.626	0.595	0.584	0.57	0.559	0.556	0.533	0.526	0.514	0.515		0.421
20	0.8	0.733	0.664	0.65	0.617	0.65	0.577	0.583	0.55	0.543	0.533	0.525	0.515	0.506	0.492	

Fig. D.1. Table des valeurs critiques pour D_{n_1, n_2} - Test de Kolmogorov Smirnov - Test bilatéral

Source : <http://www.cons-dev.org/elearning/ando/tables/tables/Table08b.html#Table08b>. Nous retrouvons d'autres tables sur le site : tests unilatéraux, tests pour $n_1 = n_2$, etc.

Dans certains cas, les tables fournissent des valeurs critiques que l'on doit comparer avec la quantité $D' = n_1 \times n_2 \times D_{n_1, n_2}$ (voir par exemple [13], tables L_I et L_{II}).

D.2 Test de Cramer - von Mises

Table de Cramer - von Mises au niveau de signification 10% pour les petits effectifs ($n_1, n_2 \leq 7$). Reproduite à partir de http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177704477 (Figure D.2)

TABLE 2 Significance Levels Near 10%					
N	M	u	$\Pr \{U \geq u\}$	t	Normalized t
4	6	472	$\frac{18}{210} = .085\ 714$.383 333	
		468	$\frac{22}{210} = .104\ 762$.366 667	
4	7	634	$\frac{32}{330} = .096\ 970$.376 623	
		631	$\frac{34}{330} = .103\ 030$.366 883	
5	6	718	$\frac{46}{462} = .099\ 567$.372 727	
		710	$\frac{48}{462} = .103\ 896$.348 485	
5	7	967	$\frac{78}{792} = .098\ 485$.371 825	
		963	$\frac{80}{792} = .101\ 010$.362 302	
6	6	1020	$\frac{43}{462} = .093\ 074$.375 000	.371 314
		1008	$\frac{59}{462} = .127\ 706$.347 222	.342 078
6	7	1374	$\frac{166}{1716} = .096\ 737$.375 458	.372 127
		1373	$\frac{172}{1716} = .100\ 233$.373 626	.370 207
		\vdots			
		1362	$\frac{194}{1716} = .113\ 054$.353 480	.349 084
		1359	$\frac{196}{1716} = .114\ 219$.347 985	.343 324
7	7	1855	$\frac{160}{1716} = .093\ 240$.382 653	.379 626
		1841	$\frac{185}{1716} = .107\ 809$.362 245	.358 330
		1827	$\frac{197}{1716} = .114\ 802$.341 837	.337 034
∞	∞		.10	.347 30	.347 30

Fig. D.2. Table des valeurs critiques à 10% pour la statistique de Cramer - von Mises

D.3 Table de Mann et Whitney

Cette table est extraite du site <http://www.cons-dev.org/elearning/stat/Tables/Tab6.html>. Elle nous fournit, pour la loi de Mann et Whitney (MW), $n_2 = 7$, $n_1 = (1, \dots, 7)$ et $U = (0, \dots, 25)$, la probabilité $P(MW \leq U)$. Nous obtenons directement la probabilité critique pour un test unilatéral, il faut la multiplier par 2 pour un test bilatéral.

n2 = 7							
n1 U	1	2	3	4	5	6	7
0	0,125	0,028	0,008	0,003	0,001	0,001	0,000
1	0,250	0,056	0,017	0,006	0,003	0,001	0,001
2	0,375	0,111	0,033	0,012	0,005	0,002	0,001
3	0,500	0,167	0,058	0,021	0,009	0,004	0,002
4	0,625	0,250	0,092	0,036	0,015	0,007	0,003
5		0,333	0,133	0,055	0,024	0,011	0,006
6		0,444	0,192	0,082	0,037	0,017	0,009
7		0,556	0,258	0,115	0,053	0,026	0,013
8			0,333	0,158	0,074	0,037	0,019
9			0,417	0,206	0,101	0,051	0,027
10			0,500	0,264	0,134	0,069	0,036
11			0,583	0,324	0,172	0,090	0,049
12				0,394	0,216	0,117	0,064
13				0,464	0,265	0,147	0,082
14				0,538	0,319	0,183	0,104
15					0,378	0,223	0,130
16					0,438	0,267	0,159
17					0,500	0,314	0,191
18					0,562	0,365	0,228
19						0,418	0,267
20						0,473	0,310
21						0,527	0,355
22							0,402
23							0,451
24							0,500
25							0,549

Fig. D.3. Table des probabilités critiques pour un test unilatéral - Mann et Whitney

D.4 Table de Kruskal et Wallis

Cette table est extraite de l'article originel des auteurs de la méthode (Table 6.1, page 617; l'article est accessible via le lien sur Wikipedia http://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance). Pour une analyse à $K = 3$ groupes, elle fournit pour chaque triplet (n_1, n_2, n_3) la probabilité critique associée à une valeur donnée de H .

D'autres tables sont accessibles sur <http://courses.wcupa.edu/rbove/eco252/252suppkey.htm>

USE OF RANKS IN ONE-CRITERION VARIANCE ANALYSIS 617

TABLE 6.1 (Continued)

Sample Sizes			H	True Probability	Approximate minus true probability		
n_1	n_2	n_3			χ^2	Γ (Linear Interp.)	B (Normal Interp.)
5	4	3	7.4449	.010	+.014	+.004	-.004
			7.3949	.011	+.014	+.004	-.004
			5.6564	.049	+.010	-.005	-.004
			5.6308	.050	+.010	-.006	-.004
			4.5487	.099	+.004	-.013	+.003
			4.5231	.103	+.001	-.016	-.000
5	4	4	7.7604	.009	+.011	+.003	-.002
			7.7440	.011	+.010	+.002	-.003
			5.6571	.049	+.010	-.004	+.000
			5.6176	.050	+.010	-.004	+.001
			4.6187	.100	-.001	-.016	+.003
			4.5527	.102	+.001	-.014	+.005
5	5	1	7.3091	.009	+.016	-.002	-.009
			6.8364	.011	+.022	+.001	-.009
			5.1273	.046	+.031	-.003	-.005
			4.9091	.053	+.032	-.002	-.002
			4.1091	.086	+.042	+.007	+.020
			4.0364	.105	+.028	-.007	+.008
5	5	2	7.3385	.010	+.016	+.004	-.004
			7.2692	.010	+.016	+.004	-.004
			5.3385	.047	+.022	+.003	+.006
			5.2462	.051	+.022	+.002	+.007
			4.6231	.097	+.002	-.018	-.005
			4.5077	.100	+.005	-.016	-.001
5	5	3	7.5780	.010	+.013	+.004	-.001
			7.5429	.010	+.013	+.004	-.002
			5.7055	.046	+.012	-.003	+.000
			5.6264	.051	+.009	-.005	-.002
			4.5451	.100	+.003	-.012	+.007
			4.5363	.102	+.002	-.014	+.005
5	5	4	7.8229	.010	+.010	+.003	-.002
			7.7914	.010	+.010	+.003	-.002
			5.6657	.049	+.010	-.003	+.001
			5.6429	.050	+.009	-.003	+.001
			4.5229	.099	+.005	-.009	+.010
			4.5200	.101	+.004	-.010	+.008
5	5	5	8.0000	.009	+.009	+.003	-.002
			7.9800	.010	+.008	+.002	-.003
			5.7800	.049	+.007	-.005	-.001
			5.6600	.051	+.008	-.004	+.001
			4.5600	.100	+.003	-.010	+.008
			4.5000	.102	+.004	-.009	+.009

Fig. D.4. Table des probabilités critiques pour le test de Kruskal et Wallis

D.5 Valeurs critiques du test de rangs signés de Wilcoxon pour échantillons appariés

Source : <http://www.cons-dev.org/elearning/stat/Tables/Tab5.html>.

N	Niveau de signification, test unilatéral		
	0,025	0,01	0,005
	Niveau de signification, test bilatéral		
	0,05	0,02	0,01
6	0		
7	2	0	
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Fig. D.5. Table des valeurs critiques pour le test de Wilcoxon - Échantillons appariés

D.6 Table des valeurs critiques de la statistique de Friedman pour la comparaison de K échantillons liés

Source : <http://www.york.ac.uk/depts/maths/tables/pdf.htm>

Upper Critical Values for the Friedman Test (k treatments and b blocks)

- Notes 1. In the table below, the critical values give significance levels as close as possible to, but not exceeding the nominal α .
2. For values of k and b beyond the range of the table below, various approximations are available.

	$k = 3$		$k = 4$		$k = 5$		$k = 6$	
b	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
2	-	-	6.000	-	7.600	8.000	9.143	9.714
3	6.000	-	7.400	9.000	8.533	10.13	9.857	11.76
4	6.500	8.000	7.800	9.600	8.800	11.20	10.29	12.71
5	6.400	8.400	7.800	9.960	8.960	11.68	10.49	13.23
6	7.000	9.000	7.600	10.20	9.067	11.87	10.57	13.62
7	7.143	8.857	7.800	10.54	9.143	12.11		
8	6.250	9.000	7.650	10.50	9.200	12.30		
9	6.222	9.556	7.667	10.73	9.244	12.44		
10	6.200	9.600	7.680	10.68				
11	6.545	9.455	7.691	10.75				
12	6.500	9.500	7.700	10.80				
13	6.615	9.385	7.800	10.85				
14	6.143	9.143	7.714	10.89				
15	6.400	8.933	7.720	10.92				
16	6.500	9.375	7.800	10.95				
17	6.118	9.294	7.800	11.05				
18	6.333	9.000	7.733	10.93				
19	6.421	9.579	7.863	11.02				
20	6.300	9.300	7.800	11.10				
21	6.095	9.238	7.800	11.06				
22	6.091	9.091	7.800	11.07				
23	6.348	9.391						
24	6.250	9.250						

Fig. D.6. Table des valeurs critiques pour le test de Friedman - Échantillons appariés

Littérature

1. S. Aïvazian, I. Enukov, L. Mechalkine, *Éléments de modélisation et traitement primaire des données*, Mir, 1986.
2. T. Bulle, *Comparaison de populations - Tests non paramétriques et analyse de variance*, Masson, 1990.
3. P. Capéraà, B. Van Cutsem, *Méthodes et modèles en statistique non paramétrique - Exposé fondamental*, Presses de l'Université de Laval et Dunod, 1988.
4. 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](http://www.dagnelie.be/extextes.html)
5. G.E. Dallal, *Nonparametric Statistics*, <http://www.jerrydallal.com/LHSP/npar.htm>
6. D. Howell, *Méthodes statistiques en sciences humaines*, De Boeck Université, 1998.
7. J.H. Klotz, *A computational Approach to statistics*, <http://pages.cs.wisc.edu/~klotz/>
8. J. McDonald, *Handbook of Biological Statistics*, <http://udel.edu/~mcdonald/statintro.html>
9. H. Motulsky, *Intuitive biostatistics*, Oxford University Press, 1995 ; Choosing a statistical test, chapter 37, <http://www.graphpad.com/www/Book/Choose.htm>.
10. R. Rakotomalala, *Comparaison de populations - Tests paramétriques*, http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf
11. G. Saporta, *Probabilités, Analyse des données et Statistique*, Dunod, 2006.
12. D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman and Hall, 2007.
13. S. Siegel, N.J. Castellan Jr., *Nonparametric statistics for the behavioral sciences*, McGraw-Hill Inc., 1988.
14. C. Wendorf, *Manuals for univariate and multivariate statistics*, <http://www.uwsp.edu/psych/cw/statmanual/index.html>
15. Wikipedia, *Non-parametric statistics*, http://en.wikipedia.org/wiki/Non-parametric_statistics