Statistique non paramétrique

Master 2

Économétrie et Statistique Appliquée

Gilbert Colletaz

22 mars 2021

Table des matières

1 Introduction **5**
- 1.0.1 Rappels . . . 5
- 1.0.2 tests paramétriques et non paramétriques . . . 6
- 1.0.3 La nature des observations : les échelles de mesure . . . 7
- 1.0.4 Les caractéristiques usuelles d'une distribution . . . 8

2 Tests sur un seul échantillon **23**
- 2.1 Test d'hypothèse sur la distribution . . . 23
  - 2.1.1 Les tests EDF : Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises . . . 24
  - 2.1.2 Le test de Shapiro-Wilk . . . 27
  - 2.1.3 Exemple 1 : test de la normalité d'un vecteur d'observations 28
  - 2.1.4 Les outils graphiques : histogramme et qqplot . . . 30
  - 2.1.5 le test de χ^2 de Pearson . . . 35
  - 2.1.6 Exemple 2 : test de normalité d'observations regroupées 39
  - 2.1.7 Exemple 3 : test de fréquences et de proportions avec proc `freq` . . . 42
  - 2.1.8 Exemple 4 : le cas de fréquences nulles . . . 44
  - 2.1.9 Probabilité critique exacte, application au χ^2 de Pearson 46
  - 2.1.10 Estimation de la probabilité critique exacte par bootstrapping . . . 50
- 2.2 Test d'hypothèse sur la valeur centrale . . . 52
  - 2.2.1 Le test du signe . . . 52
  - 2.2.2 Le test de Wilcoxon ou test des rangs signés . . . 53
  - 2.2.3 Exemple 1 : test sur la valeur de la médiane d'une série . 55
  - 2.2.4 Exemple 2 : application à un échantillon apparié . . . 56
  - 2.2.5 Les tests de McNemar et de Bowker . . . 58
  - 2.2.6 Exemple 3 : de l'utilité d'une double correction des copies 63
  - 2.2.7 Le test binomial d'une proportion . . . 64

3 Tests d'égalité de deux distributions **67**
- 3.1 Le test de la médiane . . . 67
- 3.2 Les tests de Mann-Whitney et de la somme des rangs de Wilcoxon 68
  - 3.2.1 Le test de Mann-Whitney . . . 68
  - 3.2.2 Le test de Wilcoxon . . . 70
- 3.3 Les tests EDF . . . 72
  - 3.3.1 le test de Kolmogorov-Smirnov . . . 72
  - 3.3.2 le test de Cramer-von Mises . . . 73

3.3.3 le test de Kuiper 73
3.3.4 Exemple 1 : notes à la session 1 des examens du M1 ESA selon la provenance des étudiants 74
3.3.5 Exemple 2 : le paradoxe de Simpson - analyse stratifiée et contrôle de la composition des échantillons 80

4 Tests sur plus de deux échantillons 83
4.1 Le test de Kruskal-Wallis et ses variantes 84
4.1.1 Le test de Kruskal-Wallis 84
4.1.2 Quelques variantes du test de Kruskal-Wallis 85
4.2 Test de Jonckheere-Terpstra pour alternative ordonnée 87

5 Tests d'indépendance ou d'homogénéité 91
5.1 Le χ^2 de Pearson 91
5.1.1 Test d'homogénéité d'échantillons 91
5.1.2 Test d'indépendance de variables 93
5.1.3 Exemple : L'argent fait le bonheur ! 93
5.2 Le cas des tableaux $R \times C$ 97
5.3 Le test de Fisher exact 97
5.3.1 Présentation du test pour les tables 2x2 97
5.3.2 Extension aux tables *RxC* quelconques 100
5.4 Hypothèse alternative avec tendance : le test de Cochran-Armitage 102

6 Les mesures d'association 109
6.1 le coefficient phi 109
6.2 le coefficient V de Cramer 110
6.3 Le coefficient de corrélation de Pearson 111
6.4 Le coefficient de corrélation de Spearman 112
6.5 Le coefficient de corrélation de Kendall 113
6.5.1 Illustrations des différences entre Pearson, Spearman et Kendall 114

Chapitre 1

Introduction

Ce cours a pour objectif la présentation des tests non paramétriques les plus couramment utilisés. Il se situe dans le cadre de l'inférence statistique et des tests d'hypothèse usuels : on cherche à apprécier des caractéristiques d'une population à partir d'un échantillon issu de cette population.

1.0.1 Rappels

S'agissant de savoir si les données contredisent ou ne contredisent pas une proposition, vous savez que l'on est amené à réaliser un test de signification qui nécessite la formulation d'une hypothèse nulle, H_0, et d'une hypothèse alternative, H_1.
L'hypothèse nulle porte sur la ou les valeurs d'un ou plusieurs paramètres de la population ou d'un modèle statistique. L'hypothèse alternative correspond à l'argument que l'on souhaite retenir en cas de rejet de la proposition de base. Vous avez ainsi rencontré des exemples comme :

— L'espérance d'une variable X est égale à μ_0 versus elle est différente de μ_0, soit

$$H0 : E(X) = \mu_0 \ \texttt{vs} \ H_1 : E(X) \neq \mu_0$$

— L'espérance d'une variable est la même au sein de deux populations A et B versus l'espérance au sein de A est inférieure à sa valeur au sein de B

$$H0 : \mu_A = \mu_B \ \texttt{vs} \ H_1 : \mu_A < \mu_B$$

— Au sein d'un modèle de régression donné, la variable X n'est pas une explicative pertinente

$$H0 : \beta_X = 0 \ \texttt{vs} \ H_1 : \beta_X \neq 0$$

Vous savez aussi que la conclusion est toujours formulée sur l'hypothèse nulle, soit on ne rejette pas H_0, soit on rejette H_0, et que le non rejet de H_0 ne signifie pas que la proposition qui lui est associée est vraie, mais seulement que l'information contenue dans les données ne permet pas raisonnablement de l'abandonner au profit de l'alternative.
Vous savez également que "raisonnablement" suppose que l'on ait choisi un

niveau de risque de première espèce c'est à dire une probabilité de rejeter la nulle alors qu'elle est vraie, le choix le plus courant étant une valeur, habituellement notée α prise dans le triplet $\{1\%, 5\%, 10\%\}$ sans autre justification que celle de faire comme la majorité et ayant le défaut de ne pas faire intervenir le coût associé à un rejet à tort de H_0.
Il est donc possible de contrôler le risque de première espèce en imposant une valeur pour α. En revanche on ne contrôle alors pas le risque de deuxième espèce, *i.e.* la probabilité de ne pas rejeter H_0 alors qu'elle est fausse.

Si β est cette dernière probabilité, alors $1 - \beta$ est la puissance du test, c'est à dire la probabilité de rejeter une hypothèse nulle qui serait fausse. L'idéal pour réaliser un test d'hypothèse est donc de choisir le test uniformément le plus puissant, *i.e.* celui ayant, pour tout α, la plus grande puissance.

— Exemple : supposez que dans un établissement bancaire, vous acceptiez d'accorder un prêt dès lors qu'un certain score est supérieur à 3. Pour un client i quelconque il convient donc de tester $H_0 : S_i >= 3$ versus $H1 : S_i < 3$.
  - le risque de première espèce est de ne pas accorder de crédit à un client que l'on aurait dû accepter, et donc de perdre un client sain,
  - le risque de deuxième espèce est d'accorder un crédit à un client auquel on aurait dû le refuser, et donc d'accroître le risque de son portefeuille au-delà du niveau désiré,
  - la puissance du test serait la probabilité de ne pas accorder de crédit à un client auquel on ne veut effectivement pas accorder de crédit.

Enfin, la probabilité critique où *p-value* est la probabilité que la statistique de test prenne une valeur supérieure à celle obtenue sur l'échantillon alors que H_0 est vraie. Naturellement, plus la *p-value* est proche de zéro et moins l'information tirée de l'échantillon s'accorde avec la proposition décrite par H_0. La règle de décision suivante s'impose alors :

- si *p-value*$<\alpha$, on doit rejeter H_0 en faveur de H_1
- si *p-value*$>\alpha$, on ne rejette pas H_0

Lorsque la distribution de la statistique sous H_0 est connue, les logiciels d'économétrie sont souvent en mesure d'évaluer sa *p-value*. Lorsqu'elle est inconnue, des simulations de type bootstrap permettent parfois d'approcher sa valeur. L'avantage de connaître la probabilité critique est de rendre obsolète le recours aux tables statistiques affichant les valeurs critiques de la distribution considérée pour certaines valeurs de α. Elle permet en outre de travailler avec n'importe quel seuil de risque de première espèce.

1.0.2 tests paramétriques et non paramétriques

Lorsqu'on fait l'hypothèse que les observations qui décrivent les individus sont tirées de distributions dépendant d'un nombre fini de paramètres, on parlera de tests paramétriques. Plusieurs cours de licence ou de Master 1 vous ont familiarisés avec ces derniers. Par opposition, lorsqu'on n'impose pas de distribution sur ces variables on sera dans le cadre de la statistique non paramétrique.

Certains avantages et désavantages des uns et des autres sont alors immédiatement perceptibles :

— La validité des tests paramétriques va dépendre du respect des hypothèses distributionnelles faites en amont, par exemple d'une hypothèse de normalité des observations.
— la validité des tests non paramétriques ne dépendra pas de la distribution dont sont tirées les observations. Ils seront donc naturellement plus robustes que les précédents.
— imposant plus de contraintes sur le modèle statistique, les tests paramétriques doivent dominer les tests non paramétriques lorsque les hypothèses imposées sont vraies. Comme on contrôle dans les deux cas pour le risque de première espèce, l'avantage des premiers ne peut venir que du risque de seconde espèce. En d'autres termes, si les hypothèses de distribution imposées par les tests paramétriques sont vraies, ils doivent être plus puissants que les tests non paramétriques.

Ainsi, si on ne veut pas faire d'hypothèses sur les distributions des observations, ou si la taille des échantillons est trop faible pour que le recours à des théorèmes de convergence asymptotique justifie une hypothèse de distribution particulière, on sera amené à mettre en oeuvre des tests non paramétriques.
Mais il est également possible que la nature des observations contraigne le choix du test.

1.0.3 La nature des observations : les échelles de mesure

Les statistiques, paramétriques ou non paramétriques, vont être construites sur des variables que l'on peut regrouper en deux catégories, qualitative et quantitative, elles mêmes pouvant être subdivisées selon l'échelle de mesure des observations :

1. échelle nominale : deux individus auxquels on attribue la même valeur sont supposés égaux pour un caractère étudié donné. Exemple une variable indicatrice du genre d'une personne a deux modalités valant 0 (ou 'H' ou ...) pour les hommes et 1 (ou 'F' ou ...) pour les femmes, ou encore une indicatrice de la nationalité d'individus, de la marque de voitures, etc.... Il s'agit souvent d'identifier des catégories mutuellement exclusives, *i.e.* ces variables satisfont un objectif de classification.
2. échelle ordinale : les modalités prises par la variable définissent une relation d'ordre sur la population considérée. On ne peut en particulier pas interpréter les écarts des valeurs prises par la variable en termes d'intensité : un classement ne renseigne en rien sur la distance séparant les individus classés. L'exemple type est celui des échelles de Likert que l'on rencontre fréquemment dans les questionnaires, ayant des modalités de la forme : très satisfait / satisfait / insatisfait / très insatisfait. Ces variables remplissent un objectif de hiérarchisation.
3. échelle d'intervalle : En plus de la relation d'ordre précédente on dispose d'une mesure relative à la distance séparant deux individus : contrairement à la mesure précédente, on peut comparer les écarts existant entre des observations. Ainsi si la distance entre A et B est de 4 et celle de B à C est de 2 alors on peut conclure que A est deux fois plus éloigné de B que

variables	échelle	autorise	tests
qualitatives	nominale	classification	non paramétriques
	ordinale	hiérarchisation	
quantitatives	d'intervalle	interprétation des écarts	paramétriques ou non paramétriques
	de rapport	interprétation des ratios	

TABLE 1.1 – type de données, mesures d'échelle et choix du test

B l'est de C. L'origine de ces mesures, le zéro, est fixé arbitrairement. Un exemple est donné par la mesure des températures. On ne peut pas dans ce cas interpréter en termes d'intensité le rapport existant entre deux observations : si la variation de température entre 10^0 et 11^0 celsius est la même que celle qu'il y a entre 20^0 et 21^0, en revanche il ne fait pas deux fois plus chaud lorsque l'on passe de 10 à 20 degrés celsius [1].

4. échelle de rapport (ou de ratio ou proportionnelle) : C'est une échelle d'intervalle caractérisée par l'existence d'une origine, un vrai zéro. De ce fait le rapport de deux variables définit une intensité mesurable (on peut par exemple affirmer que si deux personnes perçoivent respectivement 2000 et 2500 euros par mois alors la seconde reçoit 1.25 fois le salaire de la première [2]).

La nature des données peut contraindre l'emploi de tel ou tel test. Les tests paramétriques exigent en effet que les variables de travail soient mesurées au moins sur une échelle d'intervalle. Les tests non paramétriques sont les seuls à pouvoir être mis en oeuvre sur des variables de type nominales ou ordinales. En pratique il est donc important de connaître l'échelle de mesure des variables de travail.
Enfin, vous savez encore qu'au sein des variables quantitatives il est possible de distinguer celles à observations discrètes, comme dans le cas des dénombrements, et celles à observations continues. Naturellement si un test paramétrique suppose une distribution continue, il ne pourra pas être employé dans le premier cas.

Dans ce cours, nous serons amenés à comparer des distributions entre elles. On va donc rappeler brièvement comment sont repérées les principales caractéristiques d'une distribution.

1.0.4 Les caractéristiques usuelles d'une distribution

Si on considère que les observations contenues dans un échantillon sont des réalisations d'une même variable aléatoire, celle-ci peut être définie par

1. Pour vous persuader qu'il s'il s'agit bien d'une échelle d'intervalles, convertissons tous les chiffres de cet exemple en degrés Fahreinhet : 10^0C et 11^0C deviennent respectivement 50^0F et 51.8^0F, 20^0 et 21^0C équivalant à 68^0F et 69.8^0F. L'accroissement est dans les deux cas une constante, 1^0C ou 1.8^0F.

2. Pour illustration, les mesures des températures considérées au point précédent ne correspondent pas à une échelle de ratio puisque 10^0C et 50^0F représentent le même état de la température, de même que 20^0C et 68^0F, mais le rapport entre les mesures de ces deux états n'a pas de signification en terme de proportionalité car égal à $20/10 = 2$ si on les mesure en degrés Celsius, et égal à $68/50 = 1.36$ si on utilise les degrés Fahrenheit.

sa fonction de répartition, F. Naturellement cette fonction est la plupart du temps inconnue, mais on peut en construire une estimation $\hat{F}$. Par ailleurs, on peut vouloir préciser plusieurs éléments tels que la tendance centrale, la dispersion, la symétrie ou l'asymétrie éventuelle, le comportement de l'aléatoire aux extrémités de son support.

1. La fonction de répartition empirique
   On considère un ensemble de n réalisations $\{x_1, x_2, \ldots, x_n\}$ de variables aléatoires réelles *i.i.d.* $\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$, ayant la fonction de répartition F telle que $\forall x \in \mathbb{R}, F(x) = Pr[x_i \leq x], i = 1, 2, \ldots n$. L'estimateur naturel de F est la fonction en escalier définie pour une valeur donnée d'un réel x par :

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x) \tag{1.1}$$

   où $\mathbb{I}(x_i \leq x)$ est la fonction indicatrice définie par :

$$\mathbb{I}(x_i \leq x) = \begin{cases} 1 & \texttt{si } x_i \leq x \\ 0 & \texttt{sinon} \end{cases} \tag{1.2}$$

   - Illustration : soit les 10 réalisations suivantes : $\{8, 2, 6, 5, 3, 8, 10, 7, 1, 12\}$. Pour construire $\hat{F}$ à la main, le plus simple est évidemment d'ordonner les valeurs comme dans la table 1.2.

x	0	1	2	3	5	6	7	8	10	12	13
$\hat{F}(x)$	0/10	1/10	2/10	3/10	4/10	5/10	6/10	8/10	9/10	10/10	10/10

TABLE 1.2 – fonction de répartition empirique

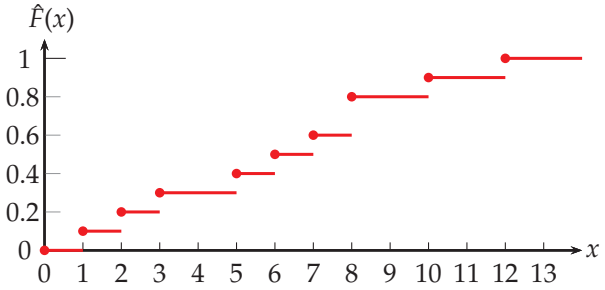


FIGURE 1.1 – graphe d'une fonction de répartition empirique

   - Propriétés de $\hat{F}$.
     Pour une valeur x donnée, $\mathbb{I}(x_i \leq x)$ est, pour tout i, une variable de Bernoulli telle que

$$Pr\left[\mathbb{I}(x_i \leq x) = 1\right] = Pr\left[x_i \leq x\right] = F(x)] \tag{1.3}$$

     et donc, $E[\mathbb{I}(x_i \leq x] = F(x)$: cette Bernoulli est un estimateur sans biais de $F(x)$. En conséquence $\hat{F}$ est également un estimateur sans biais de F :

$$E[\hat{F}] = E\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x)\right] = F(x) \tag{1.4}$$

Par ailleurs,

$$V(\mathbb{I}(x_i \leq x) = E[\{\mathbb{I}(x_i \leq x)\}^2] - E^2[\mathbb{I}(x_i \leq x)] \tag{1.5}$$
$$= 0^2[1 - F(x)] + 1^2F(x) - F^2(x) \tag{1.6}$$
$$= F(x)[1 - F(x)], \quad \texttt{et} \tag{1.7}$$
$$V[\hat{F}] = V\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(x_i \leq x)\right] \tag{1.8}$$
$$= \frac{1}{n}F(x)[1 - F(x)], \tag{1.9}$$

la dernière égalité utilisant l'hypothèse d'indépendance des n aléatoires de Bernoulli. On peut remarquer qu'asymptotiquement,

$$V[\hat{F}(x)] \underset{n\to\infty}{\to} 0, \tag{1.10}$$

et donc

$$MSE[\hat{F}(x)] = V[\hat{F}(x)] + biais[\hat{F}(x)]^2 = V[\hat{F}(x)] \tag{1.11}$$

d'où

$$\lim_{n\to\infty} MSE[\hat{F}(x)] = 0 \tag{1.12}$$

Asymptotiquement, l'erreur quadratique moyenne de $\hat{F}$ est nulle, ce qui implique la convergence en probabilité de cet estimateur :

$$\forall\epsilon > 0,\ Pr\left[\left|\hat{F}(x) - F(x)\right| > \epsilon\right] \underset{n\to\infty}{\to} 0 \tag{1.13}$$

ce que l'on note $plim(\hat{F}) = F$.

- Obtention du graphe de $\hat{F}$.

  Au sein de la `proc univariate`, on peut utiliser la commande `cdfplot`. Par défaut, en l'absence de commande de la forme `Var` *`liste de variables`*, l'appel de `cdfplot;` construit un graphique pour toutes les variables numériques de la table de travail. Lorsqu'une commande `Var` est utilisée, les variables listées dans `cdfplot` doivent être présentes dans la commande `Var=`. Ainsi les appels suivants sont valides :

```
proc univariate... ;
  cdfplot ;
```

  et

```
proc univariate... ;
  var v1 v2 v3 ;
  cdfplot v1 v3 ;
```

  En surimposition de $\hat{F}$, une option de `cdfplot` permet de représenter la fonction de répartition théorique choisie parmi onze disponibles, distributions dont les paramètres sont soit connus a priori soit, pour

certains, estimés si leur valeur n'est pas spécifiée [3]. A l'avenir, sur ces aspects, il pourrait aussi être intéressant de regarder la `proc severity` encore expérimentale dans SAS 9.4. Cette procédure considère des distributions continues soit prédéfinies, soit définies par l'utilisateur ce qui naturellement étend le champ des modélisations possibles de façon considérable par rapport à la `proc univariate`, elle permet en outre de comparer les qualités d'ajustement de plusieurs distributions sur un même jeu de données, d'afficher les quantiles associés à n'importe quelle distribution ajustée et enfin de pouvoir prendre en compte des observations censurées ou tronquées, ce que ne peut faire la `proc univariate` actuelle.

2. Les mesures de tendance centrale

   Soit une variable aléatoire X ayant une densité $f()$ et un ensemble de n réalisations $\{x_1, x_2, \ldots, x_n\}$. Les mesures de valeur centrale cherchent à résumer cet ensemble d'observations en identifiant une valeur "type". Trois mesures sont couramment utilisées : le mode, la moyenne et la médiane.

   (a) le mode : il s'agit de repérer la valeur la plus fréquente. Si X est une aléatoire discrète, c'est la valeur x du support de X pour laquelle $Pr[X = x]$ est maximale. Si X est continue, c'est la valeur de x pour laquelle $f(x)$ est maximale. Cette valeur x n'est pas nécessairement unique : on a des distributions unimodale, bimodale, trimodale, etc...Surtout utilisée pour les variables mesurées sur une échelle nominale où elle désigne la catégorie ayant l'effectif le plus élevé [4] .

   (b) la moyenne : mesure de tendance centrale la plus populaire, certainement car elle est un estimateur, sous certaines conditions, de l'espérance d'une variable aléatoire. Pour mémoire, si on a une suite d'aléatoires X_i indépendantes ayant même espérance μ_X et variance σ^2_X, alors la loi des grands nombres assure que $M_o = \frac{1}{n}\sum_{i=1}^{n} X_i$, variable évidemment centrée sur μ_X,vérifie :

   $$\forall \epsilon > 0,\ Pr\left[\left|M_o - \mu_X\right| > \epsilon\right] \underset{n\to\infty}{\to} 0 \tag{1.14}$$

   ce qui signifie que la moyenne arithmétique $M_0 = \bar{x}$, qui est un estimateur non biaisé de l'espérance converge en probabilité vers μ_X. Enfin, on montre aisément que sous les hypothèses précédentes l'écart-type de M_o est $\sigma_{M_o} = \frac{\sigma_X}{\sqrt{n}}$ [5].Cette moyenne n'a de sens que pour les observations mesurées sur au moins une échelle d'intervalle.
   La moyenne arithmétique a l'inconvénient d'être sensible aux valeurs

3. Pour le détail de ces distributions et de leurs paramètres regardez l'aide de `cdfplot` dans la `proc univariate` ainsi que la page Example 4.35 Creating a Cumulative Distribution Plot

4. Retenez que dans la table des résultats de la `proc univariate`, lorsqu'existent plusieurs modes, seul apparaît le minimum de ces valeurs modales. Une note sous la table signale toutefois la multiplicité. Ainsi, `"Note: The mode displayed is the smallest of 4 modes with a count of 3"` signale que pour la variable concernée, quatre valeurs modales sont trouvées avec des effectifs de trois individus chacune.

5. Comme vous le savez, cet écart-type est estimé par $s_{M_o} = \frac{s}{\sqrt{n}}$, avec $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$, et sous hypothèse de normalité on peut construire des IC à un seuil α pour l'espérance de la distribution, par exemple selon $\bar{x} \pm t_{(1-\frac{\alpha}{2},n-1)}s_{M_o}$ pour un IC bidirectionnel.

appel	données	données utilisées	moyenne
`proc univariate ;`	{1, 2, 3, 4, 7, 10 }	{1, 2, 3, 4, 7, 10 }	4.50
proc univariate winsorized=1 ;	{1, 2, 3, 4, 7, 10 }	{2, 2, 3, 4, 7, 7 }	4.166667
proc univariate winsorized=2 ;	{1, 2, 3, 4, 7, 10 }	{3, 3, 3, 4, 4, 4 }	3.50
proc univariate trimmed=1 ;	{1, 2, 3, 4, 7, 10 }	{., 2, 3, 4, 7, . }	4.0
proc univariate trimmed=2 ;	{1, 2, 3, 4, 7, 10 }	{., ., 3, 4, ., . }	3.50

TABLE 1.3 – Exemples de calculs de moyennes winsorisées ou tronquées

aberrantes. Lorsqu'on soupçonne la présence d'outliers, il peut être intéressant de calculer les deux autres statistiques plus robustes qui ont été proposées, toutes deux accessibles dans la `proc Univariate` :

i. la moyenne tronquée (trimmed mean) : on supprime les k premières et k dernières observations de l'échantillon et on calcule la moyenne sur les valeurs restantes. Il en résulte naturellement une statistique plus robuste que la moyenne usuelle aux petites et grandes valeurs contenues dans l'échantillon. Notez que la moyenne tronquée n'est plus une gaussienne même si au départ on dispose de réalisations normales. Cependant, si la distribution initiale est au moins symétrique, on peut construire un intervalle de confiance et faire un test de moyenne à partir de la moyenne tronquée.

ii. la moyenne winsorisée (winsorized mean) : on remplace les k plus petites valeurs par celle de rang k+1, et les k plus grandes par la valeur de rang $n - k$. En revanche, cette statistique, comme la précédente, n'est plus une gaussienne même si elle est construite à partir de réalisations gaussiennes, mais autorise également la construction d'IC et la réalisation d'un test sur l'espérance si l'échantillon initial est tiré d'une distribution symétrique.

Sous SAS, leur calcul s'effectue dans la `Proc Univariate` au moyen des options trimmed=k et winsorized=k lors de l'appel de la procédure, les deux options pouvant être simultanément présentes avec des ordres différents, selon

proc univariate data=... winsorized=k_1 trimmed=k_2 ;

la table 1.3 illustre le comportement de ces options sur le jeu d'observations $\{1, 2, 3, 4, 7, 10, \}$. Les valeurs sont présentées ici ordonnées pour faciliter la lecture ; ce n'est évidemment pas nécessairement le cas dans la table de travail.

(c) la médiane : Étant donné un échantillon constitué de n valeurs mesurées au moins sur une échelle ordinale, la médiane d'échantillon est la valeur qui sépare cet échantillon en deux parties égales. Si on note $x_{(i)}$ les valeurs initiales classées par ordre croissant alors pour n pair, M_{ed} est le centre de l'intervalle $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$, et, pour n impair, $M_{ed} = x_{(\frac{n+1}{2})}$.
Dans le cas où les observations sont regroupées en classes, on qualifie de classe médiane celle contenant le ou les rangs associés au calcul de M_{ed}. Si les observations sont mesurées au moins sur une échelle d'intervalle, on calcule la valeur médiane par interpolation linéaire

sur les bornes de cette classe médiane [6] .
La médiane d'une variable aléatoire réelle X continue de fonction de répartition F est la valeur de tout réel M_{ed} telle que $Pr[X \leq M_{ed}] \geq \frac{1}{2}$ et $Pr[X \geq M_{ed}] \geq \frac{1}{2}$. Cette valeur M_{ed} est unique si F est strictement croissante.
Par défaut la `proc univariate` calcule les quantiles à 0% (=valeur minimale de la variable), 1%, 5%, 10%, 25%(premier quartile), 50%(médiane), 75%(troisème quartile), 90%, 95%, 99%, 100%(valeur maximale de la variable). Plusieurs méthodes de calcul sont disponibles via l'option `pctldef=` et des intervalles de confiance sur ces quantiles peuvent être obtenus, soit en supposant des observations normales, avec les options `cipctlnormal` ou `ciquantnormal`, soit en l'absence d'hypothèse sur la loi des observations, avec les options `cipctldf` ou `ciquantdf` [7]

3. Les mesures de dispersion
   La mesure la plus populaire est évidemment la variance ou l'écart-type d'un ensemble d'observations. Vous savez que l'écart-type est mesuré dans la même unité que les données si ces dernières sont mesurées au moins sur une échelle d'intervalle [8]. Pour cette raison ces mesures de dispersion ne permettent souvent pas la comparaison des dispersions de plusieurs variables et, à cette fin, on leur préfère parfois le coefficient de variation défini comme la valeur absolue de l'écart-type rapporté à la moyenne des observations concernées, $CV = \frac{s}{\bar{x}}$. C'est donc un nombre pur que l'on peut interpréter comme évaluant le risque par unité de la variable [9].
   Toujours si la variable est au moins sur une échelle d'intervalle, et sur un

6. Pour exemple, soit la répartition en classes de 25 scores donnée dans le tableau ci-dessous. Pour simplifier les calculs, on fait également apparaître les effectifs cumulés :

Classe	$0-5^-$	$5-10^-$	$10-15^-$	15 `et plus`
effectif	3	8	8	6
cumul	3	11	19	25

La médiane est la note de rang 13. La classe médiane est donc celle des notes supérieures ou égales à 10 et inférieures à 15. Pour estimer la médiane, sachant que 11 personnes ont moins de 10 et que 19 ont moins de 15, on fait une interpolation sur les points (11,10) et (19,15), pour trouver l'ordonnée du point d'abscisse 13, soit

$$\frac{M_{ed}-10}{13-11} = \frac{15-10}{19-11} \Rightarrow M_{ed} = 11.25$$

7. Ces options sont placées dans la ligne d'appel de la procédure et permettent de préciser le seuil de risque à utiliser, par exemple avec `ciquantdf=(alpha=0.10)` on demande des intervalles de confiance bidirectionnels à 10% sur des observations non nécessairement gaussiennes. On peut encore construire des intervalles unidirectionnels en spécifiant la valeur du mot clef `type=`, comme dans `ciquantdf=(type=lower alpha=0.10)`, les 10% sont alors répartis en totalité à gauche. les valeurs suivantes sont autorisées : `[symmetric], lower, upper, asymmetric`. Cette dernière valeur est utile lorsque des bornes symétriques ne peuvent être construites parce qu'elles butent sur les bornes 0 ou 1. La procédure construit alors des bornes asymétriques lorsqu'elle rencontre le problème et symétriques sinon. Notez enfin que ces intervalles de confiance ne sont pas créés lorsque les données utilisées sont pondérées au moyen de l'option `weight=`.

8. avec les estimateurs OLS, non biaisé sur données gaussiennes, $s^2_{OLS} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ et l'estimateur du maximum de vraisemblance, non biaisé asymptotiquement, $s^2_{MV} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$

9. Par exemple, il serait dangereux de comparer la dispersion des températures journalières en été et en hiver au moyen des écart-types des deux séries. Pour cela, le coefficient de variation est plus adapté

plan uniquement descriptif, on peut également faire référence à l'étendue qui est simplement la différence entre les plus grandes et plus petites valeurs d'une série d'observations. L'inconvénient de cette mesure, comme d'ailleurs de l'écart-type (ou de la variance) est qu'elles sont particulièrement affectées par la présence d'outliers. En conséquence, des mesures de dispersion plus robustes on été proposées lorsqu'on soupçonne l'existence de valeurs aberrantes aux extrémités de la distribution empirique :

(a) De préférence à l'étendue, on utilise l'écart interquartile égal à la différence entre les valeurs des troisième et premier quartiles. Par définition Q3-Q1 contient la moitié des effectifs de l'échantillon. Il est naturellement moins sensible à la présence de quelques observations anormalement petites ou grandes dans l'échantillon de travail.

(b) Le coefficient de différence moyenne de Gini (*Gini Mean Difference*) : c'est la moyenne des écarts sur tous les couples d'observations qu'il est possible de construire dans l'échantillon, soit :

$$G = \binom{n}{2} \sum_{1 \le i < j \le n} \left| x_i - x_j \right| = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \left| x_i - x_j \right| \tag{1.15}$$

(c) La médiane des écarts absolus à la médiane (*Median Absolute Deviation*). Si y est le vecteur de observations,

$$MAD = M_{ed}\left(\left| y - M_{ed}(y) \right| \right), \tag{1.16}$$

où $M_{ed}(z)$ est la médiane des n observations de la variable z. Par exemple, si $y = (1, 2, 3, 4, 5)^\top$, alors $M_{ed}(y)$=3 et les écarts absolus sont $(2, 1, 0, 1, 2)^\top$, séquence dont la médiane vaut 1 et donc $MAD = 1$.

(d) Deux autres statistiques, S_n et Q_n censées pallier certaines insuffisances de *MAD* ont été proposées par Rousseeuw and Croux et sont implémentées dans la `proc univariate` [10]

Toutes ces statistiques robustes sont accessibles avec l'option `robustscale` placée dans la ligne d'appel de la `proc`.

Pour illustrer les développements qui précèdent et notamment l'obtention d'estimateurs robustes du centre et de la dispersion d'un ensemble de réalisations, nous allons utiliser la variable `score1` présente dans le fichier `Evaluation` disponible sur le site. Elle contient les valeurs d'un score établi sur 150 étudiants. Pour commencer, nous demandons des sorties standards donnant les statistiques de base, par appel sans option de la `proc univariate`. Soit le programme :

```
proc univariate data=Evaluation;
  var score1;
run;
```

La partie de l'output qui nous intéresse est reprise [11] dans la table 1.4.

10. *MAD* est inapproprié pour les distributions asymétriques. Pour plus de détails sur ces deux statistiques relativement peu usitées, voir la section *Robust Measures of Scale* dans l'aide de la `proc`

The SAS System

The UNIVARIATE Procedure
Variable: score1

Moments			
N	150	Sum Weights	150
Mean	105	Sum Observations	15750
Std Deviation	23.1226668	Variance	534.657718
Skewness	-0.5924984	Kurtosis	2.64647216
Uncorrected SS	1733414	Corrected SS	79664
Coeff Variation	22.0215874	Std Error Mean	1.88795784

Basic Statistical Measures			
Location		Variability	
Mean	105.0000	Std Deviation	23.12267
Median	106.0000	Variance	534.65772
Mode	106.0000	Range	168.00000
		Interquartile Range	26.00000

TABLE 1.4 – PROC UNIVARIATE : Les statistiques de base

Notez encore que la `proc Univariate` affiche par défaut les tables 1.5 et 1.6. La première indique les 10 valeurs les plus extrêmes de la variable `score1` ainsi que le numéro d'entrée des observations concernées. Ceci peut être utile pour le repérage rapide de valeurs aberrantes dues par exemple à des erreurs de saisie. La seconde donne les valeurs de quelques quantiles, ce qui donne une première idée de la distribution des observations.

`univariate` et les références qui y sont données.

11. Voici le détail des calculs des statistiques vues jusqu'ici : il n'y a pas d'option `weight`, en conséquence chaque observation reçoit un poids unitaire par défaut. On a 150 observations, donc :

$$\texttt{"Sum Weights"} = \sum_{i=1}^{150} 1 = 150,$$

$$\texttt{"Mean"} = \frac{\sum_{i=1}^{150} \text{score}_i}{150} = \frac{15750}{150} = 105,$$

$$\texttt{"Uncorrected SS"} = \sum_{i=1}^{150} \text{score}_i^2 = 1733414,$$

$$\texttt{"Corrected SS"} = \sum_{i=1}^{150} (\text{score}_i - \text{Mean})^2 = \sum_{i=1}^{150} \text{score}_i^2 - 150 \times 105^2 = 79664,$$

$$\texttt{"Variance"} = \frac{79664}{149} = 534.657718,$$

$$\texttt{"Std Deviation"} = \sqrt{534.657718} = 23.1226668,$$

$$\texttt{"Coeff Variation"} = \frac{23.1226668}{105} \times 100 = 22.0215874,$$

`attention dans la sortie SAS le CV est multiplié par 100,`
`il s'exprime donc en %`

$$\texttt{"Std Error Mean"} = \frac{23.1226668}{\sqrt{150}} = 1.88795784$$

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	135	151	15
34	111	152	95
51	102	153	124
55	22	157	99
63	103	169	92

TABLE 1.5 – PROC UNIVARIATE : Affichage par défaut des 10 valeurs extrêmes

Quantiles (Definition 5)	
Level	Quantile
100% Max	169.0
99%	157.0
95%	143.0
90%	130.5
75% Q3	118.0
50% Median	106.0
25% Q1	92.0
10%	79.0
5%	69.0
1%	34.0
0% Min	1.0

TABLE 1.6 – PROC UNIVARIATE : Affichage par défaut de quelques centiles

Quelques statistiques robustes sont maintenant réclamées avec le code suivant [12]

```
ods select TrimmedMeans WinsorizedMeans RobustScale;
proc univariate data=Evaluation
           trimmed=5 .05
           winsorized=.05 robustscale;
  var score1;
run;
```

Les résultats vous sont présentés dans la Table 1.7. Notez qu'on peut déduire de chacune de ces statistiques robustes de dispersion un estimateur de l'écart-type de la distribution des observations. En comparant les estimations, on peut juger de l'utilité du passage aux mesures robustes, ou au contraire, justifier l'emploi de l'estimateur standard.
Retenez aussi qu'un cas d'utilisation type de ces estimateurs robustes est naturellement la détection des *outliers* : assez souvent on les identifie comme les points dont l'écart à la moyenne est supérieur à $\pm h$ fois l'écart-type standard [13]. Sachant que ce dernier, de même que la moyenne, est sensible à la présence de ces outliers, il est préférable de repérer les valeurs s'écartant de la médiane de $\pm h$ fois l'un de ces écart-types robustes. Par exemple ici, avec une médiane de 106 et une MAD de 20.8, et en

12. Vous remarquerez que dans les options `trimmed` et `winsorized`, la valeur spécifiée peut être soit un entier, soit un pourcentage dans ce cas, le pourcentage est appliqué au nombre d'observations et k est égal à l'entier immédiatement supérieur au résultat. Par exemple, ici, $.05 \times 150 = 7.5$ et donc $k = 8$.
13. Habituellement, la valeur de h est prise dans l'ensemble $\{2.0, 2.5, 3.0\}$.

The SAS System

The UNIVARIATE Procedure
Variable: score1

Trimmed Means

Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits		DF	t for H0: Mu0=0.00	Pr > |t|
3.33	5	105.4571	1.743753	102.0094	108.9049	139	60.47713	<.0001
5.33	8	105.4254	1.739037	101.9856	108.8651	133	60.62284	<.0001

Winsorized Means

Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits		DF	t for H0: Mu0=0.00	Pr > |t|
5.33	8	105.5933	1.739734	102.1522	109.0345	133	60.69509	<.0001

Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	26.00000	19.27382
Gini's Mean Difference	25.01924	22.17272
MAD	14.00000	20.75640
Sn	20.27420	20.27420
Qn	22.21900	21.67003

TABLE 1.7 – UNIVARIATE : Exemples de sortie de statistiques robustes

retenant $h = 3$, nous identifierions comme outlier toute observation inférieure à $106-3\times20.8 = 43.6$ ou supérieure à $106+3\times20.8 = 168.4$. D'après la table 1.5, on éliminerait donc les $135^{\text{ème}}$, $111^{\text{ème}}$ et $92^{\text{ème}}$ observations. [14]

4. Asymétrie et courbure
   Tendance centrale et dispersion ne sont pas les seules caractéristiques intéressantes d'une distribution. Au-delà des moments d'ordre un et deux on peut être amené à s'interroger sur les propriétés des moments d'ordres supérieurs, notamment trois et quatre qui vont nous renseigner sur sa symétrie et sa courbure. Ces quantités sont intéressantes notamment lorsque l'on veut discuter de l'hypothèse de normalité à laquelle on se réfère souvent pour mener des tests paramétriques.
   (a) La skewness
   Il s'agit de préciser la symétrie ou la dissymétrie de la distribution. Le coefficient de skewness est défini comme le rapport du moment d'ordre 3 à la puissance troisième de son écart-type :

$$s_k = \frac{m_3}{\sigma^3} \tag{1.17}$$

14. Pour information, la procédure `boxplot`, qui produit des graphes de type *boîte à moustache* définit comme outliers les observations supérieures à $Q3+1.5\times IQR$ ou inférieures à $Q1-1.5\times IQR$ et comme *Far Outliers* celles supérieures à $Q3+3\times IQR$ ou inférieures à $Q1-3\times IQR$, où $IRQ = Q3-Q1$ est l'écart interquartile. Dans notre exemple, en utilisant les informations contenues dans les tables 1.4 et 1.6, nous aurions pour repérer les outliers une borne inférieure à $92-1.5\times26 = 53$, et supérieure à $118 + 1.5 \times 26 = 157$, et pour les *Far Outliers* des bornes à $92 - 3 \times 26 = 14$ et $118 + 3 \times 26 = 196$. Nous aurions alors un *Far Outliers*, la $135^{\text{ème}}$ observation, et il faudrait ajouter la $102^{\text{ème}}$ à la liste des outliers repérés au moyen de *MAD*.

La division de m_3 par σ^3 fait de s_k un nombre pur, qui ne dépend pas de l'unité de mesure appliquée à la variable x. Il est généralement estimé par une équation de type OLS, qui prend en compte les nombres de degrés de liberté des quantités mises en oeuvre dans le calcul de la statistique [15]. C'est en tout cas la formule par défaut de SAS, qui active implicitement une option `vardef=df` :

$$\hat{s_k} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{1.18}$$

Si on précise `vardef=n` alors il n'y a pas de correction sur les pondérations et la skewness est estimée simplement par

$$\hat{s_k} = n^{-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{1.19}$$

Lorsque la distribution est symétrique autour de l'espérance s_k vaut zéro. Il est positif pour une distribution présentant une asymétrie à droite et négatif pour une asymétrie à gauche. On utilise parfois le coefficient de skewness de Pearson défini par :

$$P_{s_k} = \frac{3(\bar{x} - M_{ed})}{s} \tag{1.20}$$

On peut montrer que ce coefficient varie entre 3 et -3 et est centré sur zéro pour une distribution symétrique. Lorsqu'il y a relativement plus de valeurs élevées dans un échantillon, la moyenne est tirée vers le haut et peut passer au-dessus de la médiane. Inversement, lorsqu'il y a relativement trop de petites valeurs, la moyenne devient inférieure à la médiane. En d'autres termes :

- La distribution est asymétrique à gauche si $\bar{x} < M_{ed}$, ou encore si P_{s_k} est négatif,
- elle est asymétrique à droite si $\bar{x} > M_{ed}$, ou encore si $P_{s_k} > 0$

Pour illustrer ces propos, on a représenté dans le graphe 1.2 la fonction de densité d'une log-normale de paramètres (1,1). Son espérance et sa médiane valent respectivement 4.48 et 2.72, son écart-type est de 5.87 , ce qui donne un coefficient de skewness $P_{sk} = \frac{3(4.48-2.72)}{5.87} = 0.90$. Cette distribution se caractérise donc par une asymétrie positive [16].

Un test de nullité de s_k a été proposé, sous l'hypothèse nulle, la quantité z définie comme

$$z = \hat{s}_k \sqrt{\frac{(n-1)(n-2)}{6n}} \tag{1.21}$$

est la réalisation d'une gaussienne centrée réduite. A ma connaissance, ce test n'est pas implanté dans SAS.

15. Les expressions données par la suite supposent un poids unitaire sur chaque observation.
16. D'ailleurs une log-normale est nécessairement dissymétrique compte tenu de son espace de définition.

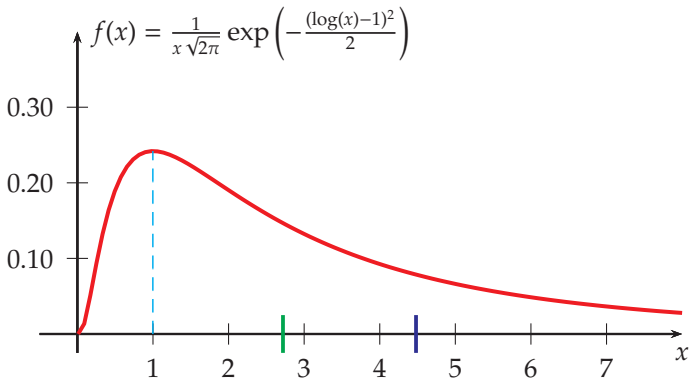


FIGURE 1.2 – densité d'une log-normale(x,1,1)
`f(x)`, `Moyenne`, `Médiane`, `Mode`

(b) La kurtosis
L'objectif est ici de caractériser l'aplatissement de la fonction de densité d'une distribution. Le coefficient d'aplatissement, ou coefficient de Pearson, est construit à partir du moment centré d'ordre 4 selon :

$$k_u = \frac{E\left[(X - E[X])^4\right]}{\sigma^4} \tag{1.22}$$

Il résulte de cette division, comme pour la skewness, un nombre pur ne dépendant pas de l'unité dans laquelle est mesurée la variable. Pour une distribution normale, le coefficient de Pearson vaut 3. Comme le plus souvent il s'agit de comparer l'aplatissement d'une distribution à celui d'une gaussienne, on met celle-ci en référence pour calculer un coefficient dit de Fisher, qui mesure l'excès de kurtosis :

$$k'_u = k_u - 3 \tag{1.23}$$

- Les distributions pour lesquelles $k'_u = 0$ sont qualifiées de mésokurtiques. C'est par construction le cas de la loi normale.
- Lorsque l'*excess kurtosis* est négatif, la distribution est qualifiée de platikurtique. C'est le cas des distributions qui présentent, relativement à la gaussienne, une plus grande concentration d'observations lorsqu'on s'éloigne de leur centre.
- Elle est qualifiée de leptokurtique s'il est positif. C'est le cas des distributions qui ont, toujours relativement à la gaussienne, un excès d'observations vers leur centre ou vers les extrémités de leur support. Le pic de leur fonction de densité est donc soit plus accentué que celui de la gaussienne et elle va alors retomber plus vite, soit moins élevé que la normale, mais leur densité va retomber moins vite. C'est par exemple le cas de la distribution de student qui est donc une distribution leptokurtique à queues épaisses [17].

17. Si $k > 4$ est le nombre de degré de liberté d'une student, alors $k'_u = \frac{6}{k-4}$.

Dans la figure 1.3, nous illustrons ces différents cas, en considérant les versions standardisées des lois hyperbolique, gaussienne et uniforme. Leurs excès de kurtosis sont respectivement de 2.0, 0.0 et -1.2. La loi hyperbolique est donc leptokurtique, la gaussienne mésokurtique et l'uniforme platikurtique.
L'estimation du coefficient de Fisher sous SAS se fait en prenant en compte les degrés de liberté, l'option `vardef=df` est activée par défaut, selon :

$$\hat{k}'_u = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{1.24}$$

Si on active l'option `vardef=n`, l'estimation est obtenue avec :

$$\hat{k}'_u = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3 \tag{1.25}$$

Un test de nullité de k'_u a été proposé : sous l'hypothèse nulle, la quantité

$$z = \hat{k}'_u \sqrt{\frac{(n-1)(n-2)(n-3)}{24n(n-1)}} \tag{1.26}$$

est la réalisation d'une gaussienne centrée-réduite. Ce test ne semble pas avoir été implanté dans SAS.

Jarque et Bera ont proposé de combiner skewness et kurtosis pour construire un test dont l'hypothèse nulle est celle d'une distribution symétrique et mésokurtique :

$$JB = n \left(\frac{\hat{k}_u^2}{24} + \frac{\hat{s}_k^2}{6} \right) \tag{1.27}$$

Sous $H0$, JB possède une distribution de χ^2 à un degré de liberté. Assez curieusement, il n'est pas implanté dans la `proc Univariate`, mais on peut le trouver dans la `proc Autoreg`.

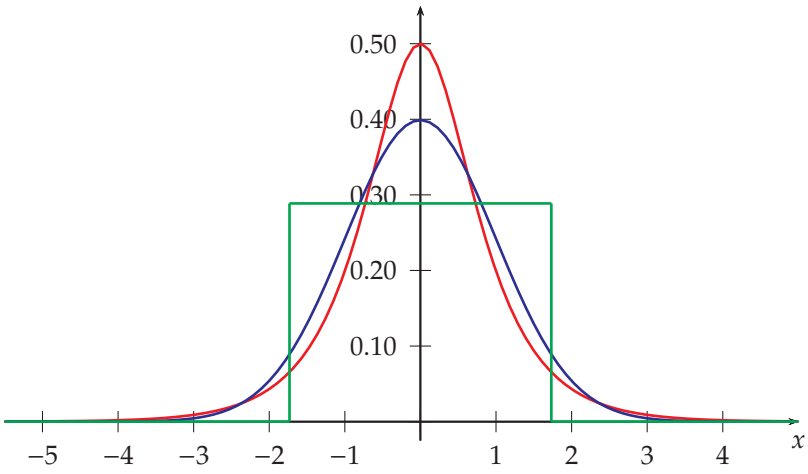


FIGURE 1.3 – densités d'une normale, d'une hyperbolique, d'une uniforme

Chapitre 2

Tests sur un seul échantillon

Dans cette partie du cours, on va supposer que l'on dispose d'un ensemble d'observations relatives à une seule variable. On suppose également qu'il s'agit de réalisations indépendantes d'une même variable aléatoire [1]. Les interrogations usuelles dans cette configuration portent soit sur la loi de probabilité de la variable, soit sur le choix d'une valeur particulière pour le centre de cette distribution.

2.1 Test d'hypothèse sur la distribution

Plusieurs raisons peuvent nous amener à vouloir réaliser un test concernant la nature de la distribution de l'aléatoire à l'origine de l'échantillon. Par exemple, pour des raisons théoriques on peut avoir un a priori sur la distribution dont seraient issues les données, et on veut vérifier que cet a priori est raisonnable. On peut aussi vouloir vérifier que les conditions d'emploi d'un test dont la validité suppose une distribution particulière ne sont pas rejetées.

C'est souvent un test de distribution gaussienne qui est effectué pour des raisons compréhensibles compte-tenu de l'omniprésence de cette hypothèse dans de nombreuses analyses, modélisations et procédures de tests. Ainsi, comme vous le savez, cette hypothèse de normalité est présente pour la dérivation de tests couramment utilisés tels que student, fisher, χ^2. Elle est présente également dans les études de corrélation, d'analyse de la variance, d'analyse en composantes principales, etc...La question de l'adéquation d'une distribution gaussienne est d'ailleurs suffisamment courante pour que dans la `proc univariate` une option `normal` suffise à déclencher l'exécution de plusieurs tests ayant comme hypothèse nulle que l'échantillon est la réalisation d'une gaussienne. Pour autant, nous pouvons aussi sous cette procédure spécifier également d'autres types de lois.

`Proc UNIVARIATE` propose plusieurs tests permettant de savoir si on rejette, au seuil de risque choisi, que les observations sont tirées dans une distribution spécifiée a priori. Trois sont fondés sur la fonction de répartition empirique (EDF tests) : Kolmogorov-Smirnov, Anderson-Darling et Cramer-von Mises. Le

1. ou de réalisations indépendantes de variables ayant la même loi.

quatrième, basé sur une logique différente et uniquement dédié à la normalité des observations est le test de Shapiro-Wilk. Par ailleurs lorsque le nombre d'observations est inférieur à 2000, un certain nombre de graphiques utiles sont également accessibles via l'option `plot` et les commandes `histogram`, `probplot`, `qqplot`. En pratique, ces graphes doivent absolument être examinés en complément des conclusions tirées des tests précédents.Enfin, nous terminerons ce chapitre avec le text de χ^2 qui qui permet d'étudier l'adéquation d'une distribution spécifiée continue ou discrète à un ensemble de données regroupées en classes.

2.1.1 Les tests EDF : Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises

Le principe de ces tests est de comparer la fonction de répartition théorique $F(x)$ spécifiée sous $H0$, et la fonction de répartition empirique, $\hat{F}(x)$ vue précédemment et définie dans l'équation (1.1).

Ces tests repose sur le théorème important suivant :

Théorème 1. *Si X est une variable aléatoire de fonction de répartition $F(x)$ inversible, alors $Y = F(X)$ est une variable uniforme sur l'intervalle* $[0, 1]$

Démonstration.

$$\begin{aligned}
\forall y \in [0,1], Pr[Y \leq y] &= Pr[F(X) \leq y], \texttt{par construction de } Y \\
&= Pr[X \leq F^{-1}(y)], \texttt{car } F() \texttt{ monotone non décroissante} \\
&= F(F^{-1}(y)), \texttt{par définition de } F() \\
&= y \ \texttt{CQFD}
\end{aligned}$$

□

Notez que ce théorème donne une solution si on veut générer des nombres x comme réalisations d'une aléatoire de fonction de répartition $F()$ presque quelconque : il suffit de générer des uniformes sur $[0, 1]$ puis de calculer $x = F^{-1}(y)$. La condition est naturellement que l'on sache calculer l'inverse de $F()$.

Le test de Kolmogorov-Smirnov

Compte-tenu des réalisations disponibles, pour statuer sur le caractère approprié de la loi de probabilité ayant une fonction de répartition $F()$, il semble naturel de regarder la distance qui sépare cette fonction théorique a priori de la fonction de répartition empirique $\hat{F}(y)$. Soit une mesure possible de cette distance,

$$D = \sup_x \left| \hat{F}(x) - F(x) \right|. \tag{2.1}$$

Sur $H0$ la fonction de répartition est $F(x)$, vous devez comprendre qu'une distance D proche de zéro est favorable à l'hypothèse nulle, et que plus D est élevée, plus on est tenté de rejeter $H0$. En pratique, cela signifie que nous devons espérer que D possède une certaine distribution, laquelle permettra de calculer des valeurs critiques permettant de juger si D est raisonnablement assez petite pour ne pas rejeter $H0$.

Dans un premier temps, on peut montrer aisément que la distribution de la statistique D, si elle existe, ne dépend pas de la fonction supposée $F()$. En effet :

$$\begin{aligned}\left|\hat{F}(x) - F(x)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[x_i \leq x\right] - F(x)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[F(x_i) \leq F(x)\right] - F(x)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[y_i \leq y\right] - y\right|\end{aligned}$$

et, dans la dernière égalité, $y_i = F(x_i)$ est la réalisation d'une uniforme sur $[0,1]$. Ainsi, $\left|\hat{F}(x) - F(x)\right| = \left|\hat{F}_{UNI}(y) - y\right|$, ou $\hat{F}_{UNI}(y)$ est la fonction de répartition empirique d'une uniforme sur $[0,1]$: pour un x donné, l'écart absolu entre $\hat{F}(x)$ et $F(x)$ ne dépend pas de $F()$. En conséquence,

$$D = \sup_x \left|\hat{F}(x) - F(x)\right| = \sup_x \left|\hat{F}_{UNI}(y) - y\right| \tag{2.2}$$

ne dépend pas de $F()$.

La seconde étape, celle de l'identification de la distribution de D est plus complexe et repose sur le théorème suivant que nous accepterons sans démonstration :

Théorème 2. *(**Théorème de Kolmogorov**) Pour un ensemble de n variables aléatoires i.i.d. de fonction de répartition continue F on a* $Pr[\sqrt{n}D \leq x] \underset{n\to\infty}{\longrightarrow} K(x)$*, où* $K(x)$ *est la fonction de répartition de Kolmogorov définie par*

$$K(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2}. \tag{2.3}$$

Pour les faibles valeurs de n il existe des tables donnant les valeurs critiques aux seuils de risque usuels [2]. Pour les tailles d'échantillons suffisamment grandes on peut utiliser les propriétés asymptotiques et donc calculer $K(x)$.

Notez enfin qu'en théorie le test de Kolmogorov-Smirnov suppose, si $F(x)$ est une fonction de répartition paramétrique, que les paramètres de cette fonction sont connus a priori. Par exemple, dans le cas d'un test de normalité, l'hypothèse nulle est formulée comme :

- $H0$: l'échantillon est un ensemble de réalisations indépendantes d'une normale d'espérance μ_0 et de variance σ_0

où μ_0 et σ_0 sont des valeurs fixées a priori par l'utilisateur du test. Lorsque les valeurs de ces paramètres sont estimées sur l'échantillon en question, alors les

2. Les premières tables ont été donnée par Smirnov, d'où le nom conjugué du test.

valeurs critiques à utiliser ne sont plus celles du test de Kolmogorov-Smirnov [3], ainsi, si on reformule $H0$ comme suit :

- $H0$: l'échantillon est un ensemble de réalisations indépendantes d'une normale d'espérance $\bar{x}$ et de variance s

où $\bar{x}$ et s sont respectivement la moyenne et l'écart-type estimés sur l'échantillon, alors le test KS se nomme test de normalité de Lilliefors et possède ses propres valeurs critiques.

Les tests d'Anderson-Darling et de Cramer-von Mises

Ces deux tests débutent avec la même logique que le test KS, à savoir examiner la distance entre la fonction de répartition théorique supposée sous $H0$ et la fonction de répartition empirique construite sur l'échantillon $\hat{F}(x)$. Ils vont différer sur deux points :

- leur distance fait intervenir l'écart quadratique, $\left(\hat{F}(x) - F(x)\right)^2$, et non plus l'écart abolu $\left|\hat{F}(x) - F(x)\right|$,
- alors que dans KS on regarde seulement la distance maximale entre les deux fonctions, ils vont considérer l'ensemble des observations.

Leur expression générale est donc :

$$Q = n \int_{-\infty}^{+\infty} \left(\hat{F}(x) - F(x)\right)^2 \Psi(x) dF(x), \tag{2.4}$$

où $\Psi(x)$ est une fonction de pondération qui va caractériser l'un ou l'autre test.

Une conséquence majeure de ces variations est que la distributions des nouvelles statistiques, et donc leurs valeurs critiques vont dépendre de la fonction $F(x)$ retenue par l'hypothèse nulle.

1. le test de Cramer-von Mises
   La fonction de pondération est $\Psi(x) = 1$ et la statistique de test

   $$W^2 = \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(x_{(i)})\right)^2 + \frac{1}{12n} \tag{2.5}$$

   où $x_{(i)}$ est la $i^{\text{ème}}$ plus petite valeur de l'échantillon.
   Une grande valeur de la statistique W^2 est un signe défavorable à $H0$: on va rejeter cette hypothèse lorsque W^2 est supérieure à sa valeur critique.
2. le test de Anderson-Darling
   La fonction de pondération est $\Psi(x) = [F(x)(1 - F(x))]^{-1}$ et la statistique

   $$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1)\log F(x_{(i)}) + (2n+1-2i)\log(1 - F(x_{(i)})\right] \tag{2.6}$$

3. Intuitivement, en utilisant des caractéristiques mesurées sur l'échantillon de travail, on intègre une information pertinente qui tend à diminuer la valeur de D par rapport à la distance qui prévaudrait si on n'avait pas eu cette information. Il faut donc ajuster les valeurs critiques à la baisse.

Rappelez-vous que dans le test de Cramer-von Mises un même poids est donné à toutes les observations quel que soit leur rang. Dans le test d'Anderson-Darling, la fonction de pondération utilisée donne plus de poids aux observations situées dans les queues de la distribution [4]. La statistique A^2 peut donc être intéressante à regarder lorsque l'utilisateur veut précisément prêter attention aux écarts entre les deux fonctions pour les valeurs situées dans les queues de la distribution [5].

2.1.2 Le test de Shapiro-Wilk

Celui-ci est construit en partie sur des statistiques d'ordre [6]. Soit m le vecteur des espérances des statistiques d'ordre de n réalisations indépendantes d'une gaussienne z centrée-réduite et V leur matrice de variance-covariance, *i.e.*

$$m = (m_1, m_2 \ldots, m_n)^\top, \text{ avec } m_i = E(z_{(i)}),\ i = 1, 2, \ldots, n \text{ et} \tag{2.7}$$

$$V_{i,j} = \text{cov}(z_{(i)}, z_{(j)}),\ i, j = 1, 2, \ldots, n \tag{2.8}$$

Soit aussi un ensemble de n réalisations $x = (x_1, x_2, \ldots, x_n)^\top$ sur lequel va porter l'hypothèse nulle $H0$: "les x_i sont des réalisations indépendantes d'une gaussienne d'espérance μ_x et de variance σ_x^2 inconnues". On peut montrer que les estimateurs *BLUE* de μ_x et σ_x^2 sont les solutions du problème [7]

$$\min\,(x - \mathbf{1}\mu_x - \sigma_x m)^\top V^{-1}(x - \mathbf{1}\mu_x - \sigma_x m), \text{ avec } \mathbf{1} = (1, 1, \ldots, 1)^\top \tag{2.9}$$

dont les solutions sont, pour des distributions symétriques et donc en particulier pour la gaussienne :

$$\hat{\mu}_x = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}, \text{ et} \tag{2.10}$$

$$\hat{\sigma}_x = \frac{m^\top V^{-1} x}{m^\top V^{-1} m} \tag{2.11}$$

l'équation 2.11 permet ainsi le calcul, du moins en théorie, et pour une gaussienne d'un estimateur de la variance inconnue σ_x^2. Pour une gausienne, le rapport de $\hat{\sigma}_x^2$ à l'estimateur sans biais usuel, $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$, doit se rapprocher de l'unité. En revanche, lorsque x n'est pas un vecteur de gaussienne alors considérer les statistiques d'ordre d'une normale standardisée pour construire $\hat{\sigma}_x$ est inapproprié : ce dernier et s_x n'estiment plus le même objet. Le rapport de $\hat{\sigma}_x^2$ à s_x^2, noté W, constitue la statistique de Shapiro-Wilk dont l'expression finale est :

$$W = \frac{\sum_{i=1}^n (a_i x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{2.12}$$

4. Vous pouvez vérifier aisément que la fonction $\Psi(x)$ est décroissante puis croissante avec $F(x)$ et atteint son minimum en $F(x) = 0.5$.

5. On sait qu'aux extrémités du support de l'aléatoire, les fonctions de répartition théorique ou empirique se rapprochent simultanément de 0 ou 1 et, en conséquence, leurs écarts tendent à disparaitrent. Pour cette raison certains préconisent l'emploi du test d'Anderson-Darling à ceux de Kolmogorov-Smirnov et Cramer-von Mises.

6. Soit $x = (x_1, x_2, \ldots, x_n)^\top$ un vecteur d'observations, on appelle statistiques d'ordre associé à x le vecteur $x_{\leq}$ constitué des observations de x reclassées par ordre croissant : $x_{\leq} = x_{(i)}, i = 1, 2, \ldots, n$, avec $x_{(i)} \leq x_{(j)}$ si $i \leq j$.

7. Sous la nulle on a évidemment $x\_i = \mu_x + \sigma z_i$

où les a_i sont un ensemble de coefficients réels donnés par

$$a = (a_1, a_2, \ldots, a_n) = \frac{m^\top V^{-1}}{\sqrt{m^\top V^{-1} V^{-1} m}} \tag{2.13}$$

et vérifiant notamment $a^\top a = 1$. Shapiro et Wilk montrent que cette statistique est, pour des distributions symétriques, indépendante de s_x^2 et de $\bar{x}$, dépendante de la taille d'échantillon n et inférieure ou égale à l'unité [8].
En pratique, pour calculer cette statistique, il est nécessaire de connaître les coefficients a_i et donc, selon l'équation 2.13, les valeurs des espérances et des covariances des statistiques d'ordre d'une gaussienne standardisée. En 1965, lors de la parution de l'article de Shapiro-Wilk, ces valeurs n'étaient connues que pour des échantillons de taille inférieure ou égale à 20 [9]. A la suite des travaux de Royston, elles sont aujourd'hui disponibles pour des tailles largement supérieures. Dans SAS 9.4, la statistique est calculée dès lors que $n \leq 2000$. Pour le calcul de la probabilité limite, une transformation dépendante de n est effectuée sur W qui renvoie une aléatoire proche d'une gaussienne standardisée sur laquelle la *p-value* est recherchée.

Pour résumer, une statistique W proche de 1 est favorable à l'hypothèse nulle de normalité des observations, alors qu'une valeur éloignée de l'unité conduit plutôt au rejet de $H0$. Notez enfin que la distribution de W est fortement asymétrique : une valeur W=0.90 peut être, selon la taille de l'échantillon, considérée comme significativement inférieure à 1 [10].

Chacun des tests précédents étant plus ou moins sensible à des déviations différentes de la normalité, on conseille de tous les effectuer afin de conforter la conclusion portant sur le statut de $H0$, en espérant évidemment qu'ils ne se contredisent pas. Pour le test de normalité, et si on veut cependant en privilégier un au regard de la puissance, alors la littérature semble indiquer que le test de Shapiro-Wilk est celui-là, viendraient ensuite Anderson-Darling, Cramer-Von Mises et enfin le test de Kolmogorov-Smirnov, y compris dans sa version Lilliefors [11].

2.1.3 Exemple 1 : test de la normalité d'un vecteur d'observations

Pour cet exemple on reprend la variable score1 déjà utilisée. La mise en oeuvre des quatre tests qui viennent d'être présentés s'effectue simplement en ajoutant l'option `normal` [12] dans la ligne d'appel de la procédure `Univariate`. Ainsi, on génère l'affichage de la table 2.1 avec le code suivant :

8. W étant indépendante de $\bar{x}$, on peut poser sans conséquence que cette moyenne est nulle et dès lors : $\sum_i (a_i y_i)^2 \leq \sum_i a_i^2 \sum_i y_i^2 = \sum_i y_i^2 \Rightarrow W \leq 1$.

9. Dans leur article, Shapiro-Wilk proposaient des approximations permettant d'afficher des valeurs critiques aux seuils de risque usuels pour des tailles d'échantillons $n \leq 50$.

10. Par exemple, Shapiro-Wilk affichent au seuil de risque de 5% une valeur critique de 0.929 pour un échantillon constitué de 30 observations. Dans ce cas, $W = 0.90$ conduirait au rejet de l'hypothèse nulle de normalité.

11. On rappelle que la distribution de la statistique D est indépendante de la distribution supposée sur les données sous $H0$ contrairement aux trois autres statistiques. Utilisant une information supplémentaire, à savoir la connaissance de la distribution sous la nulle, il semble raisonnable que celles-ci soient plus aptes à détecter les déviations que celle-là.

12. ou `normaltest`.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.967807	Pr < W	0.0014
Kolmogorov-Smirnov	D	0.061718	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.137548	Pr > W-Sq	0.0364
Anderson-Darling	A-Sq	0.82973	Pr > A-Sq	0.0331

TABLE 2.1 – UNIVARIATE : Tests de normalité sur la totalité des observations

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994575	Pr < W	0.8602
Kolmogorov-Smirnov	D	0.049052	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.046157	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.26173	Pr > A-Sq	>0.2500

TABLE 2.2 – UNIVARIATE : Tests de normalité après suppression des 3 outliers

```
proc univariate data=Evaluation normal;
  var score1;
run;
```

Selon ces premiers résultats, au seuil de 5% l'hypothèse de distribution gaussienne des scores est rejetée avec Shapiro-Wilk, Cramer-von Mises et Anderson-Darling. Seul Kolmogorov-Smirnov est favorable à $H0$. Avant de conclure que l'on est dans un cas illustrant l'absence de puissance de ce dernier, il peut être utile de considérer l'impact des observations aberrantes. Si nous supprimons les 3 observations dont les valeurs sont extérieures à l'intervalle (médiane $\pm 3MAD$), on obtient alors, sur les 147 scores restants, les résultats de la table 2.2. selon lesquels les quatre tests s'accordent pour ne pas rejeter l'hypothèse de normalité aux seuils de risque usuels. Il faut donc surtout retenir que ces statistiques de normalité sont particulièrement sensibles à la présence d'outliers puisqu'en enlevant 3 observations sur 150, on parvient à inverser les conclusions de trois d'entre eux [13].

On peut encore souligner deux autres faiblesses des quatre tests précédents :

(i) Ils manquent de puissance sur petits échantillons : lorsque n est faible, ils ont tendance à ne pas rejeter $H0$ même lorsque celle-ci est fausse,

(ii) Ils sont généralement trop puissants lorsque n est grand : de petites déviations par rapport à la normalité ont tendance à être détectées et à apparaître comme significatives,

Les conséquences peuvent être dommageables. D'une part, lorsqu'on a de petits échantillons et que l'on veut appliquer une méthode pour laquelle l'hypothèse de normalité est cruciale, ils ont tendance à ne pas détecter la non conformité à la distribution gaussienne et donc à ne pas alerter l'utilisateur. D'autre part, lorsque n est grand et qu'ils rejettent la normalité, on a souvent un théorème

13. Et surtout cet exemple rappelle qu'avant toute analyse, la première chose à faire est de nettoyer la base de données. Dans la suite de ces notes de cours, les calculs porteront sur les 147 observations restantes de la variable `score1`, après élimination des 3 outliers identifiés avec *MAD*.

central limite qui va donner de la robustesse vis-à-vis du non respect de la normalité.
Sans nier leur utilité on conçoit donc qu'ils méritent d'être complétés et dans la pratique des outils graphiques ont démontré leur intérêt.

2.1.4 Les outils graphiques : histogramme et qqplot

Deux types de graphes sont couramment utilisés [14], soit pour mettre en évidence des déviations par rapport à une distribution supposée a priori, soit, plus simplement pour révéler certaines caractéristiques des données, telles qu'asymétrie, aplatissement, multi modalités, etc...

l'histogramme

Dans la proc `Univariate` on dispose de la commande `histogram` pour afficher ce type de graphe. Elle peut être utilisée en conjonction avec la commande `var`, afin de construire un histogramme sur chacune des variables listées, comme dans

```
proc univariate;
  var x1 x2 x3;
  histogram;
run;
```

On peut aussi demander un graphe pour un sous ensemble des variables référencées par `var` comme par exemple sur les variables x3 et x2 avec

```
proc univariate;
  var x1 x2 x3;
  histogram x3 x2;
run;
```

qui donnera uniquement les sorties standards de la procédure, sans histogramme, sur x1. En l'absence de la commande `var`, des histogrammes seront construits sur les variables listées directement dans la commande `histogram`, et des sorties standards de la proc `univariate` seront générées sur toutes les variables numériques de la table SAS.
Dans notre exemple, l'histogramme sur la variable score1 représenté dans le graphe (a) de la table 2.3 est donc généré simplement par [15]

```
proc univariate;
  histogram score1;
run;
```

14. Si on se limite aux graphes offerts par la procédure `Univariate`, mais nous avons déjà précédemment signalé un autre type de graphe particulièrement utile dans l'analyse exploratoire des données, les boites à moustache obtenues avec la proc `boxplot`.

15. Avec ce code on laisse la procédure choisir le nombre d'intervalles et leur largeur. Ces deux valeurs, ainsi que d'autres caractéristiques du graphe peuvent être contrôlées grâce à un jeu d'options. Sur ces aspects, voir l'aide de la commande `histogram` de la proc `univariate`, ainsi que celle de la commande `inset` pour insérer un cartouche contenant quelques statistiques sur la variable étudiée.

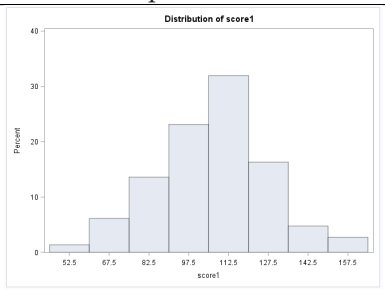
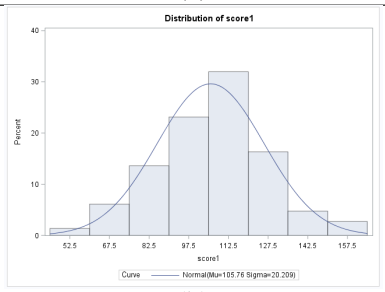
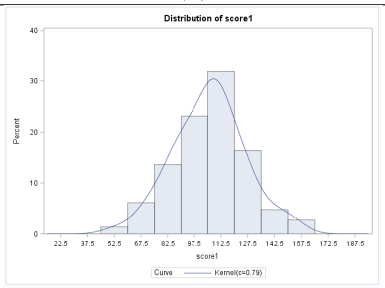
Code	Graphe associé
histogram score1 ;	 (a)
histogram score1 / normal(color=red) ;	 (b)
histogram score1 / kernel ;	 (c)

TABLE 2.3 – Proc `Univariate` - Exemples avec la commande `histogram`

Lorsque l'hypothèse de normalité est examinée, le graphe permet de voir aisément si la forme de l'histogramme est compatible avec la courbe en cloche typique de cette distribution [16], ce qui semble être le cas ici. On peut conforter cette impression en demandant, comme dans le graphe (b) de la table 2.3, le tracé de la densité d'une gaussienne. Pour cela on fait appel à l'option `normal` de la commande `histogram` :

```
proc univariate;
  histogram score1 / normal(color=red);
run;
```

Par défaut ce sont les moyenne et écart-type estimés sur l'échantillon qui sont utilisés. La ligne de code précédente est en effet équivalente à

```
  histogram score1 / normal(mu=est sigma=est color=red).
```

Il est naturellement également possible de préciser pour `mu` et `sigma` les valeurs de son choix, et même de préciser une liste de valeurs qui ferait apparaître les tracés de plusieurs fonctions de densité sur le même graphe. [17].

Enfin, plutôt que de représenter $\phi(\bar{x}, s_x) = \frac{1}{s_x\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\frac{x-\bar{x}}{s_x})^2\right)$, il est possible de demander l'affichage d'une estimation par la méthode du noyau de la fonction de densité. Le paramètre de lissage peut être imposé ou estimé par minimisation d'une approximation du critère *MISE* (Mean Integrated Square Error) avec un choix à faire parmi trois opérateurs à noyau [18] qu'il est possible de borner via des options `lower=` et `upper=`. Un exemple minimaliste est donné par

```
proc univariate;
  histogram score1 / kernel;
run;
```

avec lequel on produit la figure (c) de la table 2.3 associée à une valeur estimée $c = 0.79$. Dans notre exemple, les trois graphes (a), (b) et (c) de cette table semblent compatibles avec l'hypothèse de normalité des observations de la variable score1, confortant ainsi les résultats des tests précédents.
Retenez encore que nous avons travaillé sur l'hypothèse nulle d'une distribution gaussienne parce que c'est la plus courante, mais que d'autres distributions

16. Il faut aussi évidemment un nombre raisonnable d'observations.

17. l'argument `color=` permettrait d'attribuer une couleur particulière à la fonction de densité représentée. Comme vous le constaterez dans la figure 2.3 (b), cela n'a pas fonctionné ici.

18. Dans SAS, le paramètre de lissage est indirectement contrôlé : l'utilisateur peut spécifier la valeur d'un paramètre c qui est lié au paramètre de lissage λ par la formule :

$$\lambda = IQR \times c \times n^{-\frac{1}{5}}$$

où IQR est l'écart interquartile mesuré sur les observations. Les fonctions kernel disponibles sont la normale, la quadratique et la triangulaire définies par :

- quadratique : $K(u) = \frac{3}{4}(1 - u^2)$ avec $|u| \leq 1$,
- triangulaire : $K(u) = 1 - |u|$ et $|u| \leq 1$,
- normale : $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$, qui est l'opérateur par défaut.

Pour plus de détails sur tous ces points, Cf. votre cours ***d'économétrie non paramétrique***, et aussi, dans l'aide de la commande `histogram`, la section *Nonparametric Density Estimation Options*

peuvent être considérées, aussi bien au moyen des tests *EDF*, qu'avec la commande `histogram` [19].

le diagramme quantile-quantile

Il s'agit toujours de savoir si une collection d'observations est compatible avec une hypothèse nulle de la forme $H0$: 'les observations sont des réalisations indépendantes d'une aléatoire de fonction de répartition $F()$ spécifiée a priori'. De même que l'histogramme, un qq-plot est une aide visuelle permettant de savoir si $H0$ peut être raisonnablement maintenue et surtout de mettre en évidence les caractéristiques des observations qui sont plus en faveur d'une non acceptation de cette hypothèse nulle.

Pour le construire, on porte simplement en abscisse les quantiles de la distribution théorique supposée et en ordonnée ceux estimés au sein de l'échantillon. En cas d'égalité des quantiles théoriques et estimés, situation évidemment favorable au non rejet de $H0$, les points seront alignés sur une droite. A l'inverse plus les points s'éloigneront de cette droite de référence et plus l'utilisateur sera tenté de rejeter l'hypothèse nulle. Naturellement juger de la qualité de l'alignement sur cette seule représentation graphique laisse une part importante à l'arbitraire. Pour pallier cette difficulté, certains ont proposé de calculer la corrélation entre les deux séries de quantiles, obtenant ainsi une mesure plus objective.
L'autre aspect intéressant du *qqplot* apparaît précisément lorsque l'alignement sur la droite ne se vérifie pas. Dans ce cas, il peut être utile d'interpréter les écarts observés, ceux-ci étant susceptibles d'indiquer des caractéristiques intéressantes de la distribution des observations, telles qu'asymétrie, excès ou insuffisance de kurtosis,...
Dans la proc `univariate`, la commande réclamant le tracé d'un graphe quantile-quantile est `qqplot`. En option l'utilisateur doit préciser la distribution théorique associée à $H0$. SAS offre actuellement les distributions beta, exponentielle, gamma, log-normale, normale et de Weibull. Par défaut la gaussienne est utilisée. L'utilisateur doit aussi indiquer le nom d'une ou de plusieurs variables [20] pour la, ou lesquelles, la conformité avec la distribution spécifiée va être examinée.

Le graphique quantile-quantile de la figure 2.1 permettant l'examen de l'adéquation de la loi normale aux observations de la variable score1, est ainsi obtenu avec la ligne de code :

```
qqplot score1 / normal(mu=est sigma=est) ;
```

19. Dans cette dernière, on peut trouver une quinzaine de densités en plus de la normale avec notamment les distributions exponentielle, beta, log-normale, gamma, Weibull, qui sont aussi implémentées dans les tests *EDF*...Pour la liste complète et les mots clefs associés voir, toujours dans l'aide de la commande `histogram`, la section *Parametric Density Estimation Options*. Pour les possibilités offertes avec les tests *EDF*, voir la sous-section `Goodness-of-Fit Tests` dans la section `Details` de la proc `univariate`.

20. Comme avec `histogram`, si une commande `var` est utilisée, la variable ou la liste en question doit être présente dans cette instruction `var`. Si vous n'utilisez pas de commande `var`, un graphique quantile-quantile sera obtenu pour les variables listées dans la commande `qqplot` ainsi que les sorties standards de la proc `univariate` sur toutes les variables numériques de votre table.

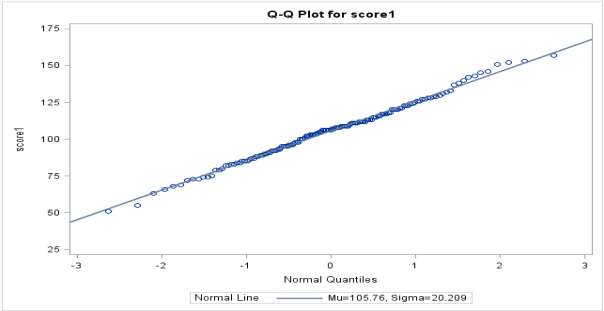


FIGURE 2.1 – Proc `Univariate` & commande : `qqplot score1;`

Dans cette figure, aucun écart du nuage de points par rapport à la droite de référence n'est suffisamment important pour mettre en doute l'hypothèse de normalité de la variable `score1` [21].

Pour mieux illustrer l'utilité de l'histogramme et du graphique quantile-quantile, nous allons considérer des cas types en générant des pseudo-réalisations d'aléatoires pour lesquelles on connaît la nature des déviations relativement à la gaussienne :

a- une variable de student à 4 degrés de liberté d'espérance nulle, leptokurtique avec un coefficient d'aplatissement de Fisher $k'_u = 3$, symétrique avec des queues de distribution épaisses,
b- une log-normale de paramètres $\mu = 0$ et $\sigma = 1$ [22], asymétrique à droite avec une skewness égale à 6.18 , leptokurtique avec $k'_u \approx 111$,
c- une uniforme sur [-1,1], symétrique et platikurtique, pour laquelle $k'_u = -1.2$,
d- une variable de χ^2 à 3 degrés de liberté, leptokurtique avec $k'_u = 3$, asymétrique à droite avec $s_k \approx 1.63$.
e- l'opposé de la log-normale précédente, qui est donc asymétrique à gauche, avec $s_k = -6.18$,
f- l'opposé de la χ^2 précédente, également asymétrique à gauche.

Nous avons créé des échantillons de 1000 points, taille relativement élevée permettant de mettre en évidence sans ambiguïté quelques caractéristiques intéressantes. Le code utilisé est le suivant :

```
data exemples;
  call streaminit(12345);
  do i=1 to 1000;
    t=rand('T',4);
    ln=rand('lognormal');
    uni=2*rand('uniform')-1;
    chi2=rand('chisquare',3);
    neg_ln = -ln;
    neg_chi2=-chi2;
```

21. Une commande telle que `qqplot score1 / normal;` aurait également fonctionné sans erreur, mais sans tracé de la droite de référence dans le rendu du graphique.
22. et donc d'espérance $\sqrt{e}$.

```
    output;
  end;
run;
proc univariate data=exemples;
  var t ln uni chi2 neg_ln neg_chi2;
  histogram t ln uni chi2 neg_ln neg_chi2 / normal;
  qqplot t ln uni chi2 neg_ln neg_chi2 / normal(mu=est
sigma=est);
run;
```

Les résultats sont présentés dans la table 2.4. On peut relever les points suivants :

1. Les queues de distribution épaisses de la student sont bien mises en évidence sur le qqplot correspondant : dans le bas de la distribution, les quantiles correspondant à une proportion donnée des observations sont atteints plus tôt dans l'échantillon qu'ils ne devraient l'être si celui-ci était un tirage de gaussienne. Par exemple, si $F_{t,4}$ et Φ sont respectivement les fonctions de répartition d'une student centrée à 4 degrés de liberté et de la gaussienne standardisée, et si on considère les fractiles à 1% et 5%. alors on a pour le premier $F^{-1}_{t,4}(1\%) \approx -3.75$ et $\Phi^{-1}(1\%) \approx -2.33$, et pour le second $F^{-1}_{t,4}(5\%) \approx -2.13$ et $\Phi^{-1}(5\%) \approx -1.64$. En conséquence, si les pseudo-réalisations reflètent raisonnablement bien les propriétés de la distribution théorique sous-jacente, le nuage de points dans cette partie du support doit se trouver sous la droite de référence. Pour les mêmes raisons, si nous regardons maintenant la partie droite de la distribution, le nuage de points de cette distribution à queues épaisses doit se trouver au-dessus de la droite de référence [23].
2. Les distributions asymétriques sont repérées dans le graphique quantile-quantile par une courbure du nuage des points, courbure à pente croissante pour les asymétriques à droite, distribution log-normale ou de χ^2, à pente décroissante pour les asymétriques à gauche, telles que les opposées des log-normales et de χ^2 des cas (e) et (f).
3. La distribution uniforme sur $[-1, 1]$ du cas (c) illustre l'impact de la troncature alors que la gaussienne est définie sur $]-\infty, +\infty[$. Ainsi, à gauche, ayant une distribution tronquée, il est obligatoire que les fractiles empiriques soient au-dessus des fractiles théoriques puisqu'avant -1, on n'a encore rencontré aucun fractile empirique. Le nuage de points débute donc nécessairement au-dessus de la droite de référence, et symétriquement, il se termine obligatoirement également en-dessous. Il ne faut évidemment pas interpréter ces caractéristiques comme étant l'indication d'une distribution à queues fines en menant un raisonnement calé sur celui tenu avec la student.

2.1.5 le test de χ^2 de Pearson

Ce test est particulièrement utile lorsque les données d'un échantillon sont réparties entre différentes classes. Lorsque seul un tel regroupement est dis-

23. Rappel, on est ici en présence de distributions symétriques et donc $\forall \alpha \in [0, 1]$, $F^{-1}_{t,4}(\alpha) = -F^{-1}_{t,4}(1-\alpha)$, et $\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$.

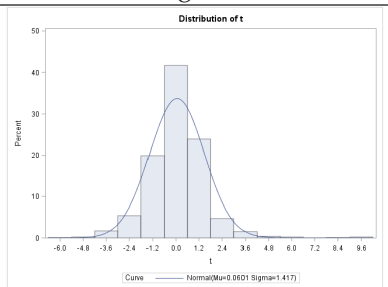
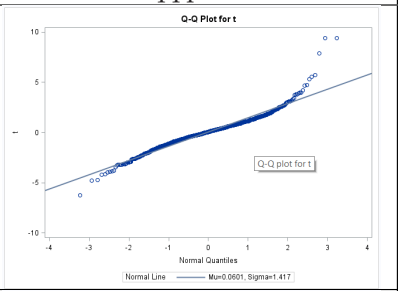
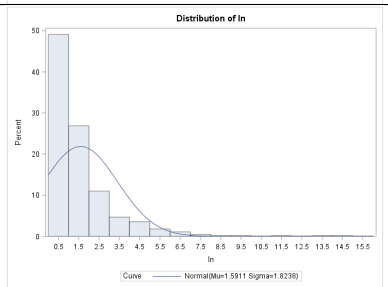
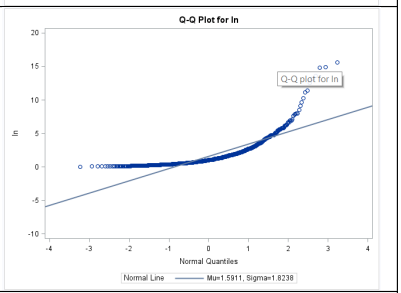
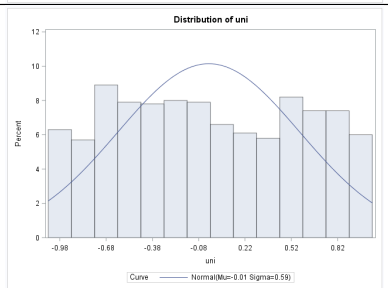
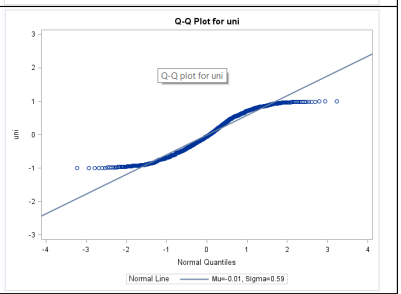
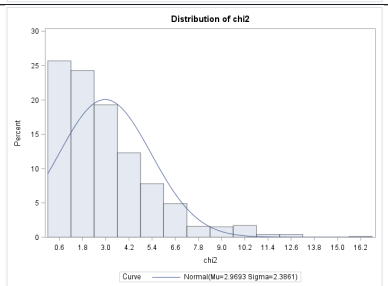
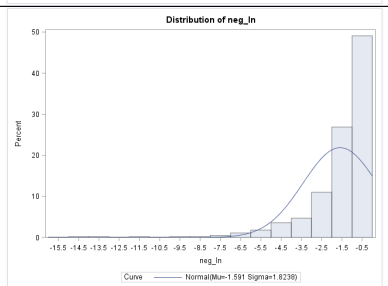
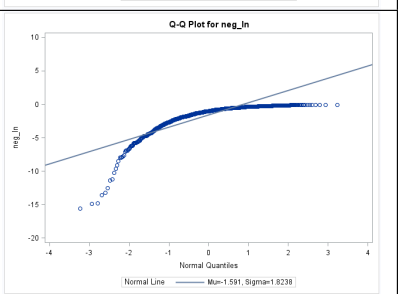
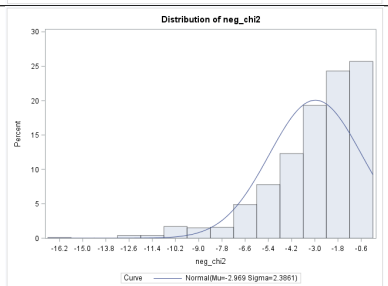
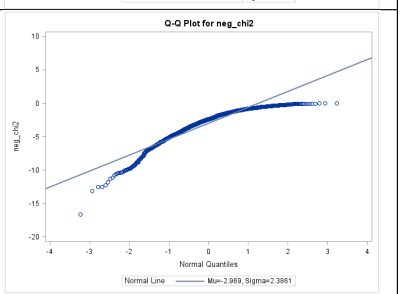
aléatoire	Histogramme	qq plot
(a) student		
(b) log-normale		
(c) uniforme		
(d) χ^2		
(e) -(lognormale)		
(f) $-\chi^2$		

TABLE 2.4 – Exemples d'histogrammes et de qqplot pour des non gaussiennes

ponible et que l'on n'a donc pas accès aux observations individuelles, les tests précédents sont inapplicables et ce test du χ^2 est alors le seul utilisable.
La logique qui le sous-tend est extrêmement simple : il s'agit de comparer les effectifs observés dans les différentes classes avec les effectifs que l'on devrait y trouver si l'hypothèse nulle était vraie. Plus les deux effectifs seront proches et moins on sera tenté de remettre en cause $H0$. A l'inverse, des écarts entre effectifs observés et espérés élevés conduiront plutôt à un rejet de $H0$. Comme d'habitude, toute la difficulté est d'apprécier l'ampleur des écarts, ce que le statisticien va traduire par : les différences entre les effectifs théoriques et observés sont-elles significativement nulles ou pas pour un seuil de risque choisi a priori ? Ce qui impose la construction d'une statistique dont on connaît la distribution sous $H0$.
Dans le cas du test de Pearson, si on est en présence d'un nombre n d'observations réparties entre c classes alors en notant O_i les effectifs observés dans la $i^{\text{ème}}$ classe et E_i les effectifs espérés dans cette même classe sous $H0$, on peut montrer que la statistique Q_P définie par :

$$Q_P = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \tag{2.14}$$

possède une distribution de χ^2 à $c-1$ degrés de liberté [24] lorsque H0 est vraie. Cependant, lorsque pour le calcul de la statistique l'utilisateur a estimé p paramètres, alors une correction doit être apportée au nombre de degrés de liberté qui devient $c-1-p$. L'obtention d'une distribution de χ^2 pour Q est fondée sur des théorèmes de convergence asymptotique supposant que les effectifs des classes soient relativement grands. En pratique, on conseille d'avoir au moins 5 observations dans chaque catégorie et, si possible, d'avoir des catégories équiprobables sous la nulle, i.e. telles que $E\_i \approx \frac{n}{c}$.

intuition pour la distribution du χ^2

On va considérer le cas le plus simple d'un découpage d'un ensemble de n observations en deux classes. Leurs effectifs respectifs seront o_1 et o_2, avec $o_1 + o_2 = n$. Soit également p_1 la probabilité d'appartenir à la première classe, et donc $p_2 = 1 - p_1$ celle d'appartenir à la seconde. Ainsi, o_1 est une somme de n variables de Bernoulli de paramètre p_1, c'est une binomiale (n, p_1) et on a $E[o_1] = np_1$, $\text{Var}[o_1] = np_1(1-p_1)$. Naturellement ici, les effectifs espérés e_1 et e_2 sont respectivement égaux à np_1 et np_2 Soit la z-transform $z = \frac{o_1 - np_1}{\sqrt{np_1(1-p_1)}}$.
Asymptotiquement la distribution de z converge vers celle d'une gaussienne centrée réduite [25] et donc, toujours asymptotiquement, $z^2 = \frac{(o_1 - np_1)^2}{np_1(1-p_1)} \underset{n\to\infty}{\to} \chi^2(1)$.

24. On se souvient que le nombre total d'observations indépendantes est un nombre n non aléatoire. En conséquence, si on considère les effectifs de c classes, ils ne peuvent pas être indépendants puisque si on connaît les effectifs de $c-1$ d'entre elles, avec n connu, l'effectif de la dernière est également donné. On perd donc un degré de liberté sur les calculs réalisés sur c classes, d'où le nombre de *dl* indiqué.

25. Comme il s'agit d'une convergence asymptotique, on conseille habituellement de vérifier que les quantités $n \times p_1$ et $n \times (1-p_1)$ soient toutes deux supérieures à 5 pour justifier le recours à la *z-transform*. Par ailleurs lorsque p_1 est proche de zéro ou de l'unité, on est dans une situation d'événement rare, et on conseille de recourir à la distribution de Poisson plutôt qu'à la gaussienne pour approximer la binomiale.

Or,

$$\begin{aligned} z^2 &= \frac{(o_1 - np_1)^2}{np_1(1-p_1)} \\ &= \frac{(o_1 - np_1)^2(1-p_1) + p_1(o_1 - np_1)^2}{np_1(1-p_1)} \\ &= \frac{(o_1 - np_1)^2}{np_1} + \frac{(o_1 - np_1)^2}{n(1-p_1)} \\ &= \frac{(o_1 - np_1)^2}{np_1} + \frac{(n - o_2 - np_1)^2}{np_2} \\ &= \frac{(o_1 - np_1)^2}{np_1} + \frac{(-o_2 + n(1-p_1))^2}{np_2} \\ &= \frac{(o_1 - np_1)^2}{np_1} + \frac{(-o_2 + np_2)^2}{np_2} \\ &= \frac{(o_1 - np_1)^2}{np_1} + \frac{(o_2 - np_2)^2}{np_2} \\ &= \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = Q_p \end{aligned} \tag{2.15}$$

Au final, on vérifie bien avec deux classes, *i.e.* lorsque $c = 2$, que la définition de la statistique de Pearson donnée par (2.14) définit bien une $\chi^2(c-1)$. La proposition énoncée pour un nombre de classes c quelconque est une généralisation de ce cas particulier.
Un autre aspect intéressant à noter ici est que le théorème central limite permet d'appliquer une distribution normalement adaptée à une variable réelle continue sur des effectifs, *i.e.* sur des réalisations d'aléatoires discrètes à valeurs entières.

La correction pour continuité

Nous venons de voir qu'il était possible d'approximer une distribution discrète à réalisations dans $\mathbb{N}$ par une distribution continue adaptée à une aléatoire à réalisations dans $\mathbb{R}$. Toutefois, un ajustement appelé "correction pour continuité" censé améliorer cette approximation a été proposé. Supposons que $\tilde{x}$ soit une aléatoire discrète à valeurs entières de densité $f(x)$ et que $\tilde{z}$ soit une aléatoire continue dans $\mathbb{R}$ de densité $g(z)$. L'approximation de $f(x)$ par $g(z)$ telle qu'utilisée au point précédent, *i.e.* sans correction, signifie que l'on approche

$$Pr[a \leq \tilde{x} \leq b] = \sum_{x=a}^{b} f(x) \tag{2.16}$$

par

$$\int_a^b g(z)dz \tag{2.17}$$

Dans le graphique 2.2 on a représenté l'histogramme correspondant à une variable aléatoire discrète à réalisations entières. Chaque rectangle a une surface

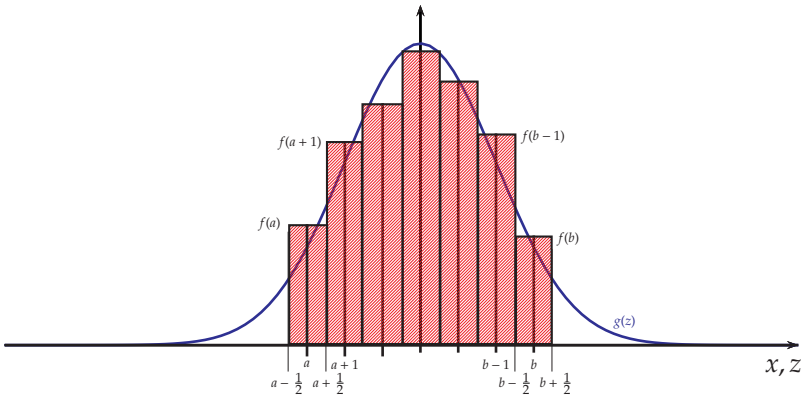


FIGURE 2.2 – densité discrète $f(x)$ et densité continue $g(z)$

égale à la probabilité de l'entier situé au centre de sa base. Ainsi, le premier rectangle, qui a comme base le segment $[a - \frac{1}{2}, a + \frac{1}{2}]$ a une surface égale à $Pr(x = a) = f(a)$, et le dernier, centré sur b a une surface égale à $Pr(x = b) = f(b)$. Ainsi la probabilité de l'équation (2.16) est donnée par la somme de tous les rectangles représentés. Maintenant, si on veut utiliser la densité de l'aléatoire continue pour approximer cette probabilité, il apparaît qu'on ne doit pas utiliser l'équation (2.17), qui donne la surface sous la courbe $g(z)$ entre a et b, mais plutôt :

$$\int_{a-\frac{1}{2}}^{b+\frac{1}{2}} g(z)dz \tag{2.18}$$

La nécessité de faire cette correction pour la continuité provient du fait que si on considère un entier c quelconque dans le support de l'aléatoire discrète $\tilde{x}$, alors $Pr[\tilde{x} = c] > 0$, en revanche avec la variable continue $\tilde{z}$ on aura toujours $Pr[\tilde{z} = c] = 0$. Pour approximer $Pr[\tilde{x} = c]$, il faudra faire $\int_{c-\frac{1}{2}}^{c+\frac{1}{2}} g(z)dz$. Pour des raisons identiques on aura :

- $Pr(x < c) = \int_{-\infty}^{c-\frac{1}{2}} g(z)dz$, et
- $Pr(x > c) = \int_{c+\frac{1}{2}}^{+\infty} g(z)dz$

Ainsi, pour une binomiale $\tilde{x}$ de paramètres (n, p), on sait que $E[\tilde{x}] = np$ et $\text{Var}[\tilde{x}] = np(1 - p)$. La correction pour continuité signifie que l'approximation de la binomiale par la gaussienne se fera selon :

$$Pr[a \leq \tilde{x} \leq b] \approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \tag{2.19}$$

2.1.6 Exemple 2 : test de normalité d'observations regroupées

Pour cet exemple on reprend la variable `score1` étudiée précédemment. Pour appliquer le test du χ^2, on va construire des bornes de classes à partir

classe	O_i	E_i	$O_i - E_i$
$]-\infty, 79.86]$	14	14.7	-0.7
$]79.86, 88.75]$	15	14.7	0.3
$]88.75, 95.16]$	16	14.7	1.3
$]95.16, 100.64]$	10	14.7	-4.7
$]100.64, 105.76]$	12	14.7	-2.7
$]105.76, 110.88]$	20	14.7	5.3
$]110.88, 116.36]$	19	14.7	4.3
$]116.36, 122.77]$	12	14.7	-2.7
$]122.77, 131.66]$	16	14.7	1.3
$]131.66, +\infty[$	13	14.7	-1.7

TABLE 2.5 – Calcul du χ^2 de Pearson sur les déciles de `score1`

des déciles de la gaussienne centrée-réduite, bornes évidemment modifiées pour s'adapter aux observations de notre variable [26]. Avec 147 observations au total, l'effectif espéré dans chaque classe construit avec les déciles est de 14.7. Les effectifs observés sont indiqués dans la table 2.5 ainsi que toutes les informations nécessaires au calcul de la statistique Q_P. Il vient :

$$\begin{aligned} Q_P &= \frac{-0.7^2}{14.7} + \frac{0.3^2}{14.7} + \frac{1.3^2}{14.7} + \frac{-4.7^2}{14.7} + \frac{-2.7^2}{14.7} + \frac{5.3^2}{14.7} + \frac{4.3^2}{14.7} + \frac{-2.7^2}{14.7} + \frac{1.3^2}{14.7} + \frac{-1.7^2}{14.7} \\ &= \frac{90.10}{14.7} \\ &= 6.12925 \end{aligned}$$

Sous $H0$: 'les observations de `score1` sont des réalisations indépendantes d'une gaussienne', cette valeur de Q serait la réalisation d'un chi^2. Deux paramètres, $\hat{\mu}$ et s ayant été estimés pour pouvoir trouver les déciles, il faut ajuster le nombre de degrés de liberté en conséquence. Ici, $c = 10$ et $p = 2$, le χ^2 est donc à 7 degrés de liberté ce qui correspond à une probabilité critique supérieure à 50%. Il n'y a donc aucune raison, aux seuils de risque usuels de rejeter cette hypothèse nulle.

On peut reproduire cet exemple avec SAS en mobilisant la procédure `Freq`. Dans le cas d'une table à entrée unique, lorsque la variable est découpée en classes, la procédure calcule par défaut le χ^2 de Pearson sous l'hypothèse nulle d'équirépartition des effectifs dans les classes. C'est précisément notre configuration, ayant utilisé les déciles pour définir les bornes des classes. Il suffira donc de définir les catégories et on utilisera pour cela la proc `format` [27]. Par la suite, dans la procédure `freq`, nous utilisons la commande `table` qui construit la table de travail avec son option `chisq` pour le calcul de Q_P. La dernière précaution

26. Rappel : pour cette variable score1, $\hat{\mu} \approx 105.76$ et $s \approx 20.21$ et donc si d est un décile de la normale standardisée, alors $\hat{\mu} + s \times d$ est le décile correspondant sur `score1`. Ainsi, les bornes des déciles constituant nos classes sont :

décile à	10%	20%	30%	40%	50%	60%	70%	80%	90%
sur gaussienne standard	-1.28	-0.84	-0.52	-0.25	0.00	0.25	0.52	0.84	1.28
sur `score1`	79.86	88.75	95.16	100.64	105.76	110.88	116.36	122.77	131.66

Pour mémoire , Φ^{-1} correspond à la fonction `Probit` de SAS.

27. Une étape `data` était évidemment également possible.

The FREQ Procedure

score1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<=79.86	14	9.52	14	9.52
79.86< <=88.75	15	10.20	29	19.73
88.75< <=95.16	16	10.88	45	30.61
95.16< <=100.64	10	6.80	55	37.41
100.64< <=105.76	12	8.16	67	45.58
105.76< <=110.88	20	13.61	87	59.18
110.88< <=116.36	19	12.93	106	72.11
116.36< <=122.77	12	8.16	118	80.27
122.77< <=131.66	16	10.88	134	91.16
>131.66	13	8.84	147	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	6.1293
DF	7
Pr > ChiSq	0.5247

FIGURE 2.3 – test de normalité de la variable `score1` au moyen du χ^2 de Pearson

consiste à demander une correction sur le nombre de degrés de liberté de la statistique, via la sous-option `df=`. sachant que la valeur par défaut serait $c - 1$. Deux possibilités sont offertes pour préciser le degré de liberté à employer : soit en termes absolus, en donnant la valeur positive visée, ici `df=7`, soit en termes relatifs, en précisant la correction négative à apporter à la valeur par défaut, ici `df=-2`. Le code pourrait donc être :

```
proc format;
  value fscore
    low-79.86 = '<=79.86'
    79.86<-88.75 = '79.86< <=88.75'
    88.75<-95.16 = '88.75< <=95.16'
    95.16<-100.64 = '95.16< <=100.64'
    100.64<-105.76 = '100.64< <=105.76'
    105.76<-110.88 = '105.76< <=110.88'
    110.88<-116.36 = '110.88< <=116.36'
    116.36<-122.77 = '116.36< <=122.77'
    122.77<-131.66 = '122.77< <=131.66'
    131.66<-high = '>131.66';
run;
proc freq data=deciles;
  tables score1 / chisq(df=-2);
  format score1 fscore.;
run;
```

L'output de ces commandes est présenté dans la figure 2.3 et on retrouve (heureusement !) le résultat des calculs faits à la main.

Composante	nb inscrits total	composition de l'échantillon
Droit-Économie-Gestion	3087	58
Lettres, Langues et Sciences Humaines	2364	32
Sciences et Techniques	3081	42
Polytech	1057	28
Instituts Universitaires de Technologie	1415	16
Observatoire des Sciences de l'Univers en région Centre	171	4
ESPE	1088	20
total	12263	200

TABLE 2.6 – répartitions des inscrits sur le site d'Orléans au 31/12/2015 et d'un échantillon de 200 étudiants pris au hasard

2.1.7 Exemple 3 : test de fréquences et de proportions avec proc `freq`

Dans cet exemple, on veut statuer sur la représentativité d'un échantillon de travail relativement à une population de référence. Cet exemple nous permettra d'introduire des tests d'égalité à une liste de fréquences ou encore à une liste de proportions.

La table 2.6 donne, au 15 décembre 2015, la répartition des inscriptions dans les diverses composantes de l'Université sur le seul site d'Orléans. Un groupe de travail, chargé de mener une étude sur les conditions de vie sur le campus a constitué un échantillon de 200 étudiants pris au hasard. Une des questions portait sur leur composante d'inscription. La répartition par composante de cet échantillon est connue et elle est également présentée dans la table 2.6. La question à laquelle nous devons répondre est de savoir si cet échantillon est représentatif de l'ensemble des étudiants classifiés selon leur seule composante d'inscription, ou si les résultats de l'étude souffrent d'un biais de sur ou de sous représentativité de telle ou telle composante ?

Disposant d'un regroupement en différentes catégories d'une population et d'un échantillon, on sait que le test du χ^2 va être adapté. La difficulté de l'exercice n'est évidemment pas théorique : de la répartition par composante de la population de référence, nous pouvons évidemment tirer les proportions théoriques de chaque catégorie, et l'application de ces proportions à l'effectif de l'échantillon permet le calcul des effectifs théoriques attendus dans chaque classe pour un échantillon de 200 personnes qui serait parfaitement représentatif. Disposant des effectifs espérés sous l'hypothèse nulle, et des effectifs observés, le calcul de la statistique Q_P est donc très simple. La table 2.7 donne ces valeurs intermédiaires à savoir, les proportions théoriques espérées pour chaque composante et les fréquences espérées pour un échantillon parfaitement représentatif.

La difficulté de cet exemple réside donc seulement dans l'écriture du code SAS permettant le calcul du χ^2 de Pearson. Déjà, la procédure `freq` permet à l'utilisateur de spécifier l'hypothèse nulle soit en indiquant les effectifs théoriques attendus, via l'option `testf=(liste de fréquences)`, soit les pourcentages attendus pour les différentes classes, via l'option `testp=(liste de proportions)`. Nous aurons donc au choix, pour un résultat sur la statistique

Composante	nb inscrits total (a)	proportion dans la population (b)=(a)/12263	composition de l'échantillon représentatif 200× (b)
DEG	3087	0.25	50
LLSH	2364	0.19	39
ST	3081	0.25	50
Polytech	1057	0.09	17
IUT	1415	0.12	23
OSUC	171	0.01	3
ESPE	1088	0.09	18
total	12263	1.00	200

TABLE 2.7 – proportions des composantes et effectifs correspondants dans un échantillon de 200 étudiants parfaitement représentatif

Q évidemment identique :

```
table composante /testf=(50 39 50 17 23 3 18);
```

ou

```
table composante /testp=(0.25 0.19 0.25 0.09 0.12 0.01 0.09);
```

La source d'erreur majeure peut provenir de la non concordance entre les chiffres de l'une ou l'autre liste, et les classes définies sur la variable : il faut absolument que 50 (ou 0.25) corresponde à l'effectif (à la proportion) théorique de la catégorie DEG, 39 (ou 0.19) à LLSH, etc. C'est l'option `order` mise dans la ligne d'appel de la procédure qui va permettre de gérer ce point. Cette option peut prendre quatre modalités :

1. `order=data` : l'ordre de création des classes par la procédure `freq` est calé sur l'ordre d'apparition des valeurs de la variable lors de la lecture de la table de données
2. `order=formatted` : l'ordre de création des classes est celui donné par les valeurs, classées par ordre croissant, précisées dans le format défini par l'utilisateur, format qui doit évidemment être appliqué à la variable étudiée.
3. `order=freq` : l'ordre de création des classes est donné par l'effectif de chaque modalité prise par la variable. Ces effectifs sont classés par ordre décroissant : la première classe correspond à la modalité qui a le plus grand effectif.
4. `order=internal` : chaque valeur différente de la variable définit une classe, l'ordre de création étant fondé sur un classement par ordre croissant de ces valeurs [28]. C'est le choix par défaut de l'option `order`.

Pour l'emploi des options `testf=` ou `testp=`, le choix de `order=` devrait être celui qui donne la maîtrise de l'ordre de création des classes à l'utilisateur, et donc `order=formatted` devrait être privilégié. Pour comprendre ce point, supposons que la création de la table 'echantillon' soit réalisée au moyen des commandes suivantes :

28. Rappel pour les valeurs alphanumériques, "a" précède "b" qui précède "c", etc.

```
data echantillon;
  input code effectif;
  cards;
700 20
500 16
600 4
100 32
400 28
200 58
300 42
;
run;
```

où `code` référencie les différentes composantes selon des valeurs définies par les services administratifs. La table 2.8 donne des exemples d'emploi de cette option `order`. Il est important, lorsqu'on utilise `testf`, de vérifier dans la table de sortie que la colonne '`Test Frequency`' donne bien les effectifs espérés sous $H0$ de chaque classe [29]

Vous devez comprendre que seul l'appel avec l'option `order=formatted` donnera des résultats qui ne seront pas affectés si, par exemple, une opération de tri est réalisée sur la table SAS contenant les données ou si on veut reproduire l'étude sur un autre échantillon. Elle facilite également les opérations de maintenance. Ainsi, en cas d'ajout ou de modification dans les codes d'identifications des groupes, les seuls ajustements à faire sont concentrés dans les commandes de la proc `format`.
Pour terminer cet exemple, on présente dans la table 2.9 la valeur du χ^2 qui est évidemment identique pour les 4 exemples de la table 2.8 [30]. Au seuil de 5% de risque on rejette l'hypothèse nulle : l'échantillon des 200 personnes ne représente pas bien la population des étudiants d'Orléans lorsque seule leur composante d'inscription est prise en compte.

2.1.8 Exemple 4 : le cas de fréquences nulles

Par défaut la `Proc Freq` élimine les modalités de fréquence nulle. Il peut arriver que l'on désire forcer leur présence. Pour cela on dispose de l'option `zeros` dans la commande `weight`.
Dans la logique de l'exemple précédent, considérons le cas suivant : trois modalités, "A", "B" et "C" d'une variable X dont les effectifs totaux seraient respectivement 40, 5 et 55. Dans ce cas, il est aisé de voir qu'un échantillon de 10

29. Avec l'emploi de `testp`, la même vérification faite sur les proportions espérées sous $H0$ s'impose avec la colonne alors intitulée '`Test Percent`'. Ainsi, dans le dernier exemple, l'appel de la procédure `freq` aurait donné les mêmes résultats avec :

```
proc freq data=echantillon order=formatted;
  table code / testp=(0.25 0.09 0.12 0.19 0.01 0.09 0.25) chisq;
  weight effectif;
  format code fcomposante.;
run;
```

et la colonne `Test Percent` doit reproduire dans un ordre identique la liste des pourcentages spécifiée par `testp=`.

30. Avec l'option `testp=`, on obtient une statistique de χ^2 légèrement différente, égale à 13.9518, en raison des approximations numériques, qui ne modifie en rien la conclusion de cet exemple.

```
proc freq data=echantillon order=data;
  table code / chisq
  testf=(18 23 3 39 17 50 50);
  weight effectif;
run;
```

code	Frequency	Test Frequency
700	20	18
500	16	23
600	4	3
100	32	39
400	28	17
200	58	50
300	42	50

```
proc freq data=echantillon order=freq;
  table code / chisq
  testf=(50 50 39 17 18 23 3);
  weight effectif;
run;
```

code	Frequency	Test Frequency
200	58	50
300	42	50
100	32	39
400	28	17
700	20	18
500	16	23
600	4	3

```
proc freq data=echantillon;
  table code / chisq
  testf=(39 50 50 17 23 3 18);
  weight effectif;
run;
```

code	Frequency	Test Frequency
100	32	39
200	58	50
300	42	50
400	28	17
500	16	23
600	4	3
700	20	18

```
proc format;
value fcomposante
    100 = 'Lettres, Langues et Sciences Humaines'
    200 = 'Droit, Économie, Gestion'
    300 = 'Sciences et Techniques'
    400 = 'Polytech'
    500 = 'Instituts Universitaires de Technologie'
    600 = "Observatoire des Sciences de l'Univers en région Centre"
    700 = 'ESPE';
run;
proc freq data=echantillon order=formatted;
  table code / chisq
  testf=(50 18 23 39 3 17 50);
  weight effectif;
  format code fcomposante.;
run;
```

code	Frequency	Test Frequency
Droit, Économie, Gestion	58	50
ESPE	20	18
Instituts Universitaires de Technologie	16	23
Lettres, Langues et Sciences Humaines	32	39
Observatoire des Sciences de l'Univers en région Centre	4	3
Polytech	28	17
Sciences et Techniques	42	50

TABLE 2.8 – Exemples d'utilisation des options `order` et `testf` ou `testp` dans la procédure `freq`

Chi-Square Test for Specified Frequencies	
Chi-Square	13.6200
DF	6
Pr > ChiSq	0.0342

TABLE 2.9 – Exemple 3 : test de représentativité d'un échantillon

individus parfaitement représentatif doit être constitué de 4 individus de type "A", 0.5 de type "B" et 5.5 de type "C". Supposons qu'un échantillon de travail de 10 individus ait été constitué et qu'il soit composé de 4, 0 et 6 individus de type "A", "B" et "C". Pour juger de la représentativité de cet échantillon, on exécuterait les commandes suivantes :

```
data exemple;
input x $ echantillon;
cards;
A 4
B 0
C 6
run;
proc freq data=exemple order=data;
tables x / testf=(4 0.5 5.5) chisq;
weight echantillon;
run;
```

Avec cette syntaxe, le test de Chi-2 n'est pas réalisé, et dans le journal apparaît la remarque ERROR : The number of TESTF values does not equal the number of levels. For the table of x, there are 2 levels and 3 TESTF values. L'origine du problème vient de ce que la modalité "B" dont la fréquence est nulle dans l'échantillon n'est pas retenue dans le tableau créé par la `proc freq` dans lequel ne subsiste donc que deux modalités ce qui est incompatible avec l'indication des trois fréquences théoriques dans `testf=`. Pour rétablir la situation, il faut donc forcer la présence de la modalité "B" dans le tableau de la `procfreq`, ce que l'on fait en utilisant l'option `zeros` de `weight`. Ainsi, avec

```
weight echantillon / zeros;
run;
```

Les calculs réclamés sont bien effectués, et on obtient ici une statistique égale à 0.5455.

2.1.9 Probabilité critique exacte, application au χ^2 de Pearson

Toutes les statistiques vues jusqu'ici se caractérisent par une convergence seulement asymptotique vers une distribution connue de laquelle on peut extraire des valeurs critiques ou calculer une probabilité critique. Ces résultats naturellement très utiles ne sont cependant pas entièrement satisfaisants car il arrive que l'on doute de l'utilité des théorèmes de convergence asymptotique. En premier lieu, la taille des échantillons de travail peut être insuffisante pour y faire référence, par ailleurs, la vitesse de convergence peut dépendre des caractéristiques de l'échantillon telles qu'asymétrie ou aplatissement de la distribution des observations de sorte que la "qualité" d'un test peut varier sensiblement simplement à la suite d'un changement d'échantillon. Pour ces raisons, on préfère lorsque cela est possible, calculer une probabilité critique exacte.
La procédure `freq` permet le calcul de la probabilité critique exacte de plusieurs statistiques dont le Q_P de Pearson. Dans ce dernier cas, le cadre d'analyse est

assez simple et, pour faciliter la compréhension du concept de probabilité critique exacte et surtout des calculs à réaliser pour en trouver la valeur, nous allons rester sur ce test.

Par définition, la probabilité critique exacte associée à une statistique est la probabilité que se réalise, sous $H0$ un événement au moins aussi extrême que celui observé sur l'échantillon de travail. Le repérage de ces événements au moins aussi extrêmes va souvent dépendre de la formulation alternative $H1$: avec un test unidirectionnel, les événements extrêmes sont recherchés dans une seule direction. S'il est bidirectionnel, les déviations relativement à $H0$ sont à considérer dans plusieurs directions. De ce point de vue, la statistique de Pearson est unidirectionnelle : la valeur du chi^2, évidemment positive, dépend de la distance entre des effectifs observés et des effectifs espérés sous $H0$ et est telle que plus cette distance est grande, plus la statistique Q_P est élevée.
Pour résumer, nous sommes en présence d'une table donnant la répartition de n individus au sein de c classes, où n, la taille de l'échantillon est une constante. Sur cette table, et pour une certaine hypothèse nulle, nous pouvons calculer les effectifs espérés pour chacune des classes et finalement la valeur de la statistique sur l'échantillon de travail, soit Q_{P_0}. Il est alors possible de construire toutes les tables répartissant n individus entre c classes, et d'évaluer sur chacune la statistique de Pearson qui lui correspond. On considérera alors comme événements au moins aussi extrêmes que celui observé sur l'échantillon, les répartitions ayant des statistiques supérieures ou égales à Q_{P_0}. La somme des probabilités de ces événements est par définition égale à la probabilité critique exacte. Il reste alors à savoir calculer la probabilité de survenue d'une table quelconque, ce que l'on peut faire au moyen de la distribution multinomiale.
La distribution multinomiale est une généralisation de la binomiale. Cette dernière considère les résultats de n expériences indépendantes dans un univers de résultats à deux états possibles de probabilités respectives p et $1-p$. Avec la multinomiale, nous avons k états possibles pour les résultats, avec un ensemble de probabilités $p_1, p_2, \ldots, p_k$ vérifiant $\sum_{i=1}^{k} p_i = 1$. Soit après n expériences les nombres o_i, de réalisations du $i^{\text{ème}}$ état, avec $i = 1 \ldots, k$. Par exemple, si nous répartissons n individus indépendants entre k classes exhaustives, alors o_1 est l'effectif de la première classe, p_1 la probabilité d'appartenir à cette première classe, etc...La distribution multinomiale donne la probabilité de survenue du k-uple $(o_1, o_2, \ldots, o_k)$ connaissant l'effectif total n et le système de probabilités $p_1, p_2, \ldots, p_k$:

$$Pr[o_1, o_2, \ldots, o_k] = \frac{n!}{o_1! o_2! \ldots o_k!} p_1^{o_1} p_2^{o_2} \ldots p_k^{o_k} \tag{2.20}$$

On dispose alors de tous les éléments permettant le calcul de la probabilité critique exacte d'un χ^2 de Pearson. Pour illustrer la démarche, considérons une répartition entre deux classes, et 4 individus à répartir. Supposons que l'échantillon de travail place trois personnes dans la première classe et une dans la deuxième. Le vecteur des observations est donc $o = (o_1, o_2) = (3, 1)$ et $n = 4$. Posons $H0$:"les deux classes sont équiprobables", *i.e.* $p_1 = p_2 = 1/2$ sous $H0$. Toujours sous cette hypothèse, les effectifs espérés constituent le vecteur $e = (e_1, e_2) = (np_1, np_2) = (2, 2)$. La statistique de Pearson pour cet échantillon vaut :

$$Q_{P_0} = \frac{(3-2)^2}{2} + \frac{(1-2)^2}{2} = 1$$

tables possibles (o_1, o_2)	Q_P	Pr[table]
(4,0)	$\frac{(4-2)^2}{2} + \frac{(0-2)^2}{2} = 4$	$\frac{4!}{4!0!}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = 0.0625$
(3,1)	$\frac{(3-2)^2}{2} + \frac{(1-2)^2}{2} = 1$	$\frac{4!}{3!1!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 = 0.25$
(2,2)	$\frac{(2-2)^2}{2} + \frac{(2-2)^2}{2} = 0$	$\frac{4!}{2!2!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = 0.375$
(1,3)	$\frac{(1-2)^2}{2} + \frac{(3-2)^2}{2} = 1$	$\frac{4!}{1!3!}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = 0.25$
(0,4)	$\frac{(0-2)^2}{2} + \frac{(4-2)^2}{2} = 4$	$\frac{4!}{0!4!}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 = 0.0625$

TABLE 2.10 – Exemple de calcul d'une probabilité critique exacte sur un χ^2 de Pearson

et on sait qu'asymptotiquement cette statistique serait sous $H0$ la réalisation d'un χ^2 à 1 degré de liberté. Il est par ailleurs évident que dans cette configuration, le recours aux propriétés asymptotiques est particulièrement douteux. Pour obtenir la probabilité critique exacte il faut créer toutes les tables qui répartissent 4 individus entre 2 classes, évaluer la statistique Q_P de chacune et additionner les probabilités des tables pour lesquelles Q_P est supérieure ou égale à Q_{P_0}. La table 2.10 détaille ces calculs. Quatre tables vérifient $Q_P \geq Q_{P_0}$, et la probabilité critique exacte du test de Pearson est donc égale à la somme des quatre probabilités correspondantes, soit 62.5%.

On peut reproduire cet exemple avec SAS. La commande `exact statistique` réclame le calcul de la probabilité exacte de la statistique dont le nom est précisé. Sachant cela, il suffit d'exécuter les lignes suivantes :

```
data exact;
  input x pond;
  cards;
1 3
0 1
;
run;
proc freq data=exact;
  table x / chisq;
  exact chisq;
  weight pond;
run;
```

Pour récupérer en sortie les tables présentées dans la figure 2.4, tables conformes aux calculs réalisés à la main. Notez que dans cet exemple, nous avons deux catégories équiprobables, i.e., $p_1 = p_2 = 0.5$, ce qui est, comme on le sait, la configuration par défaut retenue par proc freq.
Dans le cas général, le système de probabilité de la multinomiale n'a pas de raison de supposer l'égalité de ces probabilités entre elles : avec k classes, et si p_i est la probabilité d'appartenir à la $i^{\text{ème}}$, alors les seules conditions sont $p_i \geq 0, i = 1, 2 \ldots, k$, et $\sum_{i=1}^{k} p_i = 1$. On doit donc être en mesure de préciser les valeurs du vecteur des k probabilités en vigueur sous l'hypothèse nulle et ce

code	Frequency	Test Frequency
Droit, Économie, Gestion	58	50
ESPE	20	18
Instituts Universitaires de Technologie	16	23
Lettres, Langues et Sciences Humaines	32	39
Observatoire des Sciences de l'Univers en région Centre	4	3
Polytech	28	17
Sciences et Techniques	42	50

TABLE 2.11 – Compositions observée et sous l'hypothèse nulle d'un échantillon de 200 étudiants

sont naturellement les options `testp` ou `testf` qui vont être utilisées [31]. Afin d'illustrer ce point, nous allons reprendre la question de la représentativité d'un échantillon de 200 étudiants vue dans la sous-section précédente. La table 2.11 rappelle la composition de cet échantillon selon la composante d'inscription des individus interrogés ainsi que les fréquences afférentes à un échantillon de 200 personnes qui serait parfaitement représentatif de la population des étudiants d'Orléans au regard de leurs filières d'étude. C'est naturellement la colonne intitulée "TEST Frequency" qui contient les effectifs espérés sous la nulle.

Il suffit donc d'ajouter la commande `exact chisq` dans le programme qui réclamait le calcul de la statistique de Pearson pour obtenir le résultat recherché. L'appel à la procédure `freq` pourrait ainsi être :

```
proc freq data=echantillon order=formatted;
  table code / chisq testf=(50 18 23 39 3 17 50);
  exact chisq;
  weight effectif;
  format code fcomposante.;
run;
```

On va se garder cependant de lancer ce programme, à moins d'être prêt à attendre quelques jours pour lire le résultat final. La masse de calculs qu'il réclame est en effet trop grande pour un ordinateur personnel de puissance usuelle. La difficulté provient du nombre de répartitions possibles de n individus dans k classes, qui est égal à $\frac{(n+k-1)!}{(k-1)!n!}$, soit ici, avec $n = 200$ et $k = 7$ un total de 9.8619410^{10} tableaux possibles. Cet exemple inabouti illustre l'inconvénient majeur des calculs exacts à savoir le temps de calcul nécessaire à leur complétion qui devient rapidement rédibhitoire , sauf sur de petits échantillons [32].
Compte-tenu de cette difficulté, on peut se demander si, plutôt que de calculer la probabilité critique exacte, il n'est pas possible de l'estimer avec une bonne précision et des temps de calculs raisonnables. Une solution est fournie par la technique du bootstrap.

31. On se souvient que la taille de l'échantillon est une constante n, on peut donc logiquement soit donner directement le vecteur des probabilités, via `testp`=$(p_1, p_2, \ldots, p_k)$, soit le donner indirectement en précisant les effectifs espérés sous $H0$ avec `testp`=$(e_1, e_2, \ldots, e_k)$, la procédure se chargeant du calcul $p_i = e_i/n, i = 1, 2, \ldots, k$.

32. Les utilisateurs raisonnables n'utiliseront d'ailleurs pas la commande `exact` sans son option `maxtime=` qui donne en secondes le temps accordés à son exécution. Ainsi, avec `exact chisq / maxtime=30;` si les calculs ne sont pas terminés au bout de 30 secondes, les objets non encore évalués seront déclarés comme valeur manquante dans la sortie, et l'exécution du programme se continuera, avec un message d'alerte dans le journal.

The FREQ Procedure

x	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1	25.00	1	25.00
1	3	75.00	4	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	1.0000
DF	1
Asymptotic Pr > ChiSq	0.3173
Exact Pr >= ChiSq	0.6250

WARNING: The table cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test.

FIGURE 2.4 – Reproduction de l'exemple du calcul d'une valeur critique exacte dans proc `freq`

2.1.10 Estimation de la probabilité critique exacte par bootstrapping

Plutôt que de chercher à reconstruire la totalité des tableaux répartissant n individus entre k classes, la méthode de Monte Carlo va en générer un certain nombre au hasard, tous de taille n, et la démarche explicitée précédemment va être reprise sur ces tableaux, *i.e.* calcul de la statistique d'intérêt et comptabilisation du nombre de tableaux pour lequel la statistique est supérieure ou égale à Q_{P_0} , ce sont des répartitions au moins aussi extrêmes que celle effectivement observée sur l'échantillon de travail ; puis estimation de la probabilité critique exacte, $\hat{p}_{MC}$ par le rapport entre cet effectif des événements extrêmes et le nombre total de tableaux simulés. La technique est mise en oeuvre grâce à l'option `MC` de la commande `exact`. L'utilisateur peut également préciser le nombre d'échantillons bootstrappés à construire, via `N=`, et faire en sorte que leur construction soit reproductible au moyen de l'option `seed=` . Enfin, il n'y a pas de règle donnant le nombre d'échantillons à construire, mais on peut, pour une valeur de N donnée, évaluer un intervalle de confiance sur l'estimation de la probabilité exacte et donc avoir une idée de la précision atteinte. Pour cela, on utilise le fait que $\hat{p}_{MC}$ est une proportion et a donc un écart-type asymptotique $s_{\hat{p}_{MC}} = \sqrt{\hat{p}_{MC}(1-\hat{p}_{MC})/(N-1)}$. L'intervalle affiché est finalement donné par $\hat{p}_{MC} \pm \left(z_{\alpha/2 \times s_{\hat{p}_{MC}}}\right)$, ou $z_{\alpha/2}$ est le centile de la gaussienne standardisée correspondant au seuil de risque α. Ce seuil α est lui-même gouverné par l'option `ALPHA=` de la commande `exact avec la valeur par défaut de 1%, qui construit alors les bornes d'un intervalle à 99% de confiance.`.

On aura par exemple :

exact chisq / mc ;

qui demande la création de 10000 échantillons bootstrap, valeur par défaut de l'option `N=` qui ne pourront pas être reproduits, et affichage de bornes d'un intervalle de confiance à 99%, ou encore :

exact chisq / N=20000 ;

commande qui implicitement invoque `MC` et la génération de 20000 échantillons non reproductibles, ou :

Chi-Square Test for Specified Frequencies	
Chi-Square	13.6200
DF	6
Asymptotic Pr > ChiSq	0.0342

Monte Carlo Estimate for the Exact Test	
Pr >= ChiSq	0.0358
95% Lower Conf Limit	0.0332
95% Upper Conf Limit	0.0383
Number of Samples	20000
Initial Seed	235128001

TABLE 2.12 – Estimation de la probabilité critique exacte par méthode de Monte Carlo

exact chisq / seed=123 ;

qui intègre implicitement l'option `MC` et réclame l'estimation de la probabilité exacte sur le χ^2 de Pearson au moyen 10000 échantillons créés au hasard, avec des résultats qui pourront être reproduits avec l'emploi de la même valeur pour l'option `seed=`, ou enfin :

exact chisq / alpha=0.05 ;

qui demande un intervalle à 95% sur l'estimation de la probabilité critique exacte, à partir de 10000 échantillons bootstrap non reproductibles [33].

On termine cette partie en demandant une estimation de la probabilité exacte sur l'échantillon de 200 étudiants considéré dans la sous-section précédente :

```
proc freq data=echantillon order=formatted;
  table code / chisq
  testf=(50 18 23 39 3 17 50);
  exact chisq / n=20000 alpha=0.05;
  weight effectif;
  format code fcomposante.;
run;
```

Les résultats obtenus peuvent être lus dans la table 2.12. On remarquera que pour cet exemple, l'approximation asymptotique est satisfaisante, et l'estimation de la probabilité critique exacte ne fait que confirmer la mauvaise représentativité de l'échantillon [34].

33. Plus généralement, si $0 < \alpha < 1$ est la valeur précisée dans la commande, l'intervalle de confiance résultant est de $100(1-\alpha)\%$.

34. Dans le log de SAS apparaît également le message :

```
NOTE: PROCEDURE FREQ used (Total process time):
real time     0.39 seconds
cpu time      0.14 seconds
```

qui rappelle le gain en termes de temps d'exécution de l'option `MC` comparativement aux heures que réclame la commande `exact`.

2.2 Test d'hypothèse sur la valeur centrale

On dispose d'un échantillon constitué de n réalisations indépendantes $x_1, \ldots, x_n$ d'une variable aléatoire x et on s'interroge sur la valeur centrale de la distribution de cette aléatoire. Alors que les tests paramétriques utilisent l'espérance comme mesure de la valeur centrale et son estimateur usuel qu'est la moyenne empirique, les tests non paramétriques vont travailler sur la médiane.
Dans ce cadre, les deux tests non paramétriques que nous allons présenter, test du signe et test de Wilcoxon, sont les plus utilisés. Ils sont implémentés dans la proc `univariate`, où ils sont d'ailleurs associés au test de student sur la moyenne d'un échantillon, qui est également la statistique paramétrique la plus populaire lorsqu'on veut réaliser un test sur la valeur d'une espérance.

2.2.1 Le test du signe

Les hypothèses considérées sont de la forme :

— $H0$: la médiane de la distribution est égale à M_0, versus
— $H1$: la médiane de la distribution est différente de M_0

où M_0 est une valeur constante fixée a priori. Ce jeu d'hypothèses a une écriture équivalente

— $H0 : Pr[x_i \geq M_0] = p = \frac{1}{2}$, versus
— $H1 : Pr[x_i \geq M_0] = p \neq \frac{1}{2}$

Le test du signe consiste à remplacer les observations plus grandes ou égales à M_0 par un signe '+' et celles qui lui sont inférieures par un signe '-'. Si l'hypothèse nulle est vraie alors le nombre de signes '+', soit n^+, doit être 'proche' du nombre de signes '-', noté n^-. Un nombre n^+ très largement supérieur à n^- est défavorable à $H0$ et tend à signaler que la valeur de M_0 retenue est peut être trop faible. Si n^+ est faible relativement à n^-, cela est encore défavorable à l'hypothèse nulle et indique que la valeur M_0 pourrait être révisée à la baisse. Indépendamment de la loi de la variable x, ces deux nombres, n^+ et n^-, sont sous $H0$ des sommes de réalisations de variables de Bernoulli de paramètre $\frac{1}{2}$ et donc des binomiales de paramètres n et $\frac{1}{2}$. Par convention la statistique de test est le nombre n^+, et on a en conséquence :

$$Pr[n^+ = n_1] = \frac{n!}{n_1!(n-n_1)!} 0.5^{n_1} 0.5^{n-n_1}, \text{ et} \tag{2.21}$$

$$F[n_1] = Pr[n^+ \leq n_1] = \sum_{j=0}^{n_1} \frac{n!}{j!(n-j)!} 0.5^j 0.5^{n-j}$$

$$= \sum_{j=0}^{n_1} \frac{n!}{j!(n-j)!} 0.5^n \tag{2.22}$$

Ainsi,

— Lorsque l'alternative est unidirectionnelle de la forme $H1$: "La médiane est supérieure à M_0" , c'est donc une trop forte valeur de $n+$ qui incitera au rejet de $H0$ en faveur de $H1$. Dans ce cas, la région de rejet pour un seuil de risque α sera de la forme $\{n^+|n^+ \geq n_\alpha\}$, ou n_α est le plus petit entier tel que $\sum_{j=n_\alpha}^{n} \frac{n!}{j!(n-j)!} 0.5^n \leq \alpha$.
— Si l'alternative est de sens opposé, soit $H1$: "La médiane est inférieure à M_0", c'est un nombre n^+ trop faible qui jouera en défaveur de $H0$ et la région de rejet sera $\{n^+|n^+ \leq n'_\alpha\}$, n'_α étant alors le plus grand entier positif tel que $\sum_{j=0}^{n'_\alpha} \frac{n!}{j!(n-j)!} 0.5^n \leq \alpha$. Remarquez que la loi binomiale étant symétrique lorsque $p = 1/2$, on a $n'_\alpha = 1 - n_\alpha$.
— Enfin pour une hypothèse alternative bidirectionnelle, $H1$: "La médiane est inférieure à M_0", on rejettera au seuil de risque α lorsque n^+ sera extérieur à l'intervalle $]1 - n_{\alpha/2}, n_{\alpha/2}[$.

Dans la procédure `univariate`, SAS affiche non pas la statistique usuelle n^+, mais la statistique $M = (n^+ - n^-)/2$, où n^+ et n^- sont maintenant les effectifs des valeurs de l'échantillon respectivement strictement supérieures et strictement inférieures à la médiane supposée M_0. Le calcul de M exclut les valeurs égales à M_0, et donc ici $(n^+ + n^-) \leq n$. Enfin, la procédure n'affiche pas les valeurs critiques, mais directement la probabilité critique exacte sous $H0$ de la statistique M égale à $\sum_{j=0}^{\min(n^+,n^-)} \frac{(n^+ + n^-)!}{j!(n^+ + n^- - j)!} 0.5^{(n^+ + n^- - 1)}$

2.2.2 Le test de Wilcoxon ou test des rangs signés

Le test du signe est donc construit sur des variables de Bernoulli associées à l'hypothèse $H0$: "la médiane est égale à M_0". De cette proposition découle en effet deux événements équiprobables sous la nulle : chaque valeur de l'échantillon observé est soit inférieure, soit supérieure à $M0$. On peut noter alors que l'ampleur des écarts des observations à la médiane n'est pas considérée : seul le signe importe pour le décompte des '+' et des '-'. On peut alors espérer gagner en puissance en intégrant l'information relative à l'amplitude des écarts à M_0, et c'est ce que va faire le test de Wilcoxon.

Ce test pourra être employé sur des observations indépendantes mesurées au moins sur une échelle d'intervalle. La démarche est la suivante : on commence par construire les écarts des observations à la médiane supposée $e_i = x_i - M_0, i = 1, 2, \ldots, n$. On ordonne ensuite par ordre croissant les valeurs absolues de ces écarts, et on conserve pour chaque écart son rang, $r(|e_i|$ et son signe : $s_i = 1$ si $e_i > 0$ et $s_i = 0$ sinon. Cette aléatoire s est une variable de Bernoulli qui vérifie, si l'hypothèse nulle est vraie, $E(s_i) = 1/2$ et $\sigma^2_{s_i} = 0.25$. Si on construit la variable SR^+ comme somme des rangs des écarts positifs :

$$SR^+ = \sum_{i=1}^{n} s_i r(|e_i| \tag{2.23}$$

alors, en supposant l'absence d'observations de même rang [35], il vient :

$$E(SR^+) = 0.5 \sum_{i=1}^{n} r(|e_i| = 0.5 \sum_{i=1}^{n} i = \frac{n(n+1)}{4}, \text{ et} \tag{2.24}$$

$$Var(SR^+) = \sum_{i=1}^{n} i^2/2 - (i/2) = n(n+1)(2n+1)/24 \tag{2.25}$$

Si $SR^- = \sum_{i=1}^{n}(1 - s_i)r(|e_i|$ est la somme des rangs des écarts négatifs, alors $SR^+ + SR^- = \frac{n(n+1)}{2}$ [36], et le fait d'observer $SR^+ \gg SR^-$ est une indication favorable à l'hypothèse d'une médiane supérieure à M_0. La statistique de Wilcoxon, ou statistique des rangs signés, est habituellement définie comme la plus petite des deux valeurs SR^+ et SR^-.
On peut remarquer que dans la construction de la statistique, les écarts négatifs et positifs reçoivent le même poids, ce qui se comprend dans le cas d'une distribution symétrique. Ainsi l'hypothèse nulle de la statistique de Wilcoxon doit s'énoncer comme $H0$:" La médiane de l'aléatoire pour laquelle on dispose de n réalisations indépendantes est M_0 **et** sa distribution est symétrique" : c'est donc un test joint, et il faut en tenir compte, notamment en cas de rejet.
Lorsque n est supérieure à 25, on admet qu'une *z-transform* sur $SR = \min(SR^+, SR^-)$ est une approximation raisonnable d'une gaussienne centrée-réduite, permettant ainsi aisément de trouver des valeurs critiques, ou de fournir une estimation de la probabilité critique. La variable de test devient alors, en utilisant les équations 2.24 et 2.25 :

$$z = \frac{SR - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}} \tag{2.26}$$

La statistique de Wilcoxon calculée par la procedure `univariate` apporte des améliorations par rapport à notre présentation sur plusieurs points :

- Lorsque le nombre d'observations est inférieur ou égal à 20, elle calcule la probabilité critique exacte de la statistique,
- elle tient compte des observations ayant le même rang :
  - dans le calcul de la statistique elle même qui est évaluée selon :

$$S = SR^+ - \frac{n_t(n_t+1)}{4} \tag{2.27}$$

    où n_t est le nombre d'observations disponibles après que les e_i nuls aient été éliminés des calculs, et ou on a attribué aux e_i de même rang leur rang moyen [37],
  - dans l'expression de la variance de la *z-transform* lorsque le nombre d'observations est supérieur à 20, la probabilité critique affichée est celle d'un student à $n_t - 1$ degrés de liberté définit par

$$S\sqrt{\frac{n_t - 1}{n_t V - S^2}}, \text{ avec} \tag{2.28}$$

35. En particulier les observations pour lesquelles l'écart est nul sont purement et simplement retirées des calculs et non prises en compte dans le total des observations disponibles, n.
36. Puisqu'alors $SR^+ + SR^-$ est simplement la somme des n premiers rangs ou, dit autrement, la somme des n premiers entiers naturels.
37. De sorte que la somme des rangs n'est pas affectée par la manipulation.

Tests for Location: Mu0=10				
Test	Statistic		p Value	
Student's t	t	-0.55108	Pr > |t|	0.5903
Sign	M	-3	Pr >= |M|	0.1460
Signed Rank	S	-8	Pr >= |S|	0.5532

TABLE 2.13 – Test de l'hypothèse d'une valeur centrale de la distribution de durées d'attente égale à 10 minutes

$$V = \frac{1}{24} n_t(n_t + 1)(2n_t + 1) - \frac{1}{48} \sum t_i(t_i + 1)(t_i - 1) \tag{2.29}$$

où la somme est faite sur les groupes d'écarts égaux en valeur absolue et ou t_i est l'effectif du i^{ème} groupe.

2.2.3 Exemple 1 : test sur la valeur de la médiane d'une série

Le test du signe et le test de Wilcoxon sont réalisés dans la procédure `univariate`. Il suffit dans la ligne d'appel de cette proc de préciser la valeur supposée a priori de la médiane dans l'option `MU0=` , et naturellement le ou les noms des variables concernées dans la commande `VAR`. Dans cet exemple, nous disposons de 15 temps d'attente d'usagers dans un service public, le temps d'attente est la durée en minutes séparant l'instant où une personne prend un ticket indiquant un numéro d'appel et le moment où il est est invité à se présenter à un guichet. La question est de savoir si la valeur centrale de la distribution des durées d'attente peut être de 10 minutes.

```
data attente;
  input duree @@;
  cards;
9 5 6 10 10 6 8 3 15 6 7 17 8 20 10
;
run;
proc univariate data=attente mu0=10;
  var duree;
run;
```

On récupère en sortie la table 2.13, laquelle ne permet pas, aux seuils de risque usuels, de rejeter l'hypothèse nulle que ce soit à l'aide d'un test paramétrique de student [38], qui suppose la normalité des observations et évalue la valeur centrale par l'espérance, ou par les deux tests non paramétriques qui évaluent la valeur centrale par la médiane et ne font pas d'hypothèse particulière sur la distribution des observations, si ce n'est celle de symétrie pour le test de Wilcoxon.

38. Pour information, la moyenne et l'écart-type de la moyenne des 15 durées enregistrées dans l'échantillon sont respectivement proches de 9.33 et 1.210. La statistique de student affichée est $t = -0.55 \approx \frac{9.33-10}{1.21}$ et est sous $H0$ la réalisation d'une student à 14 degrés de liberté.

2.2.4 Exemple 2 : application à un échantillon apparié

Le cas d'un échantillon apparié est rencontré assez souvent. Par exemple, on dispose d'un premier score sur un ensemble d'individus, puis, à la suite d'une intervention, le score est réévalué sur les mêmes individus. La question posée est celle de l'efficacité de l'intervention, qui peut être un contact commercial, une opération de promotion, une action publicitaire, etc...L'inefficaité se traduisant par une reproduction à l'identique des scores dans la seconde mesure : dans ce cas, le vecteur des écarts des deux scores est nul. On qualifie également d'échantillon apparié la configuration où le deuxième échantillon n'est pas nécessairement composé des mêmes individus que le premier, mais d'individus ayant des distributions identiques sur un certain nombre de variables de contrôle telles que l'âge, le sexe, le statut familial, professions et catégories socio-professionnelles, etc... Les personnes constituant les deux échantillons ne sont donc pas nécessairement indépendantes. On exige cependant toujours qu'au sein d'un même échantillon les individus le constituant soient indépendants entre eux. Lorsque des paires d'individus semblables entre les échantillons peuvent être créées, et cela est évidemment plus simple lorsque l'appariement concerne les mêmes individus, alors on peut juger de l'efficacité de l'intervention en créant pour chaque item l'écart de ses scores pour tester la nullité de la valeur centrale de la distribution de ces écarts. Techniquement, à l'exception d'une étape `data` nécessaire à la création des différences, on retrouve à l'identique l'exemple précédent.
Pour illustrer la démarche, nous allons considérer deux séquences de notes. La variable `note1` contient les évaluations obtenues à une première épreuve par 40 étudiants, et `note2` celles obtenues par les mêmes étudiants à la seconde épreuve, après une séance de correction de la première. On va admettre que le niveau de difficulté des deux examens est identique, et la question est donc celle de l'efficacité de la séance de correction : a-t'elle été profitable aux étudiants ? Dans l'affirmative, la valeur centrale de la variable d'écart devrait être positive pour chaque étudiant. Il s'agit de vérifier cela sur les réalisations `note2-note1`. Une valeur nulle signalerait l'inefficacité de la séance de correction, et une valeur négative qu'un seuil de saturation sur la matière concernée a été atteint. Les commandes pourraient être les suivantes, sachant que la variable `id` contient l'identifiant propre à un étudiant ce qui a permis de s'assurer de l'exactitude des couples formés dans l'appariement.

```
data notes;
  input id exam1 exam2 @@;
  cards;
2016236   8   9
2016956   7.5   7
2016267   9   9.5
2016355   15.5   15
2016915   12   13
2016324   13   13
2015110   6.5   7
2014195   5   6
2016254   12   14
2016175   9   8.5
```

```
2016192   10.5   11
2016451   11   11
2014125   .50   0
2015337   3.5   6
2016859   14.5   14
2016405   12   13.5
2016754   8.5   8
2016632   11   10.5
2016741   10.5   11.5
2016795   9   10
2016235   11.5   10
2016204   10   11.5
2016210   14   16
2016563   7.5   9
2016621   10.5   10.5
2016384   11   9.5
2016429   8.5   11.5
2016951   12   13.5
2015386   4.5   5.5
2016739   9.5   11
2016754   12   11
2016452   10   9.5
2016492   8   8.5
2015923   8   9
2016125   7.5   10
2016548   12.5   11
2016835   10.5   11
20166978   9   9.5
2016327   5   5
2016897   10   11;
run;
data notes;
  set notes;
  ecart=exam2-exam1;
run;
proc univariate data=notes mu0=0;
  var ecart;
run;
```

La procédure affiche alors la table 2.14. Avec un seuil de risque de 5%, les tests de student, du signe et de Wilcoxon permettent le rejet de l'hypothèse nulle en faveur d'une moyenne et d'une médiane des écarts strictement positives [39].

39. On se rappelle que la probabilité critique reportée dans la table 2.14 correspond à une hypothèse alternative bidirectionnelle. Il faut donc la diviser par deux avant de la comparer à son seuil de risque si on donne à $H1$ une formulation unidirectionnelle, ici de la forme "la médiane, ou la moyenne est supérieure (ou inférieure) à zéro" : on est alors bien sur des valeurs inférieures à 5% . Quant au signe de l'alternative les valeurs positives des 3 statistiques signalent qu'elles sortent de leur intervalle de confiance en transperçant la borne haute : l'alternative n'a de sens qu'avec une formulation $H1$: "la médiane, ou l'espérance, est supérieure à 0".

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.963019	Pr > |t|	0.0052
Sign	M	6	Pr >= |M|	0.0652
Signed Rank	S	171.5	Pr >= |S|	0.0048

TABLE 2.14 – Test de l'hypothèse de nullité de la valeur centrale des écarts sur échantillon apparié

2.2.5 Les tests de McNemar et de Bowker

Le test de McNemar

Souvent utilisé pour statuer sur d'éventuelles modifications qui seraient survenues à la suite d'un apport d'information ou d'une action nouvelle. En théorie, ce n'est rien d'autre qu'un test du signe sur échantillon apparié tel que vu précédemment. Cependant, alors que nous avions accès aux informations concernant chacun des individus constituant l'échantillon apparié, dans le test de McNemar les données sont regroupées sous forme d'un tableau (2×2).

Pour présenter ce test, nous allons considérer une configuration simple : on veut étudier l'impact d'un débat sur les changements d'opinion des électeurs. Pour cela, on dispose d'un échantillon de personnes initialement réparties en deux sous-groupes, l'un préférant le candidat A, l'autre le candidat B. A la suite du débat entre A et B on reconstruit la répartition entre les deux sous groupes A et B. Pour chaque individu, quatre configurations sont donc envisageables :

- $(A \to A)$ un individu initialement classé A est toujours dans le sous-groupe A à la fin de l'expérience,
- $(B \to B)$ un individu initialement classé B est toujours dans le sous-groupe B à la fin de l'expérience,
- $(A \to B)$ un individu initialement classé A a modifié son opinion suite au débat et est passé dans le sous-groupe B,
- $(B \to A)$ un individu initialement classé B a modifié son opinion et est passé dans le sous-groupe A à la fin de l'expérience.

Les exemples d'applications sont évidemment nombreux : choix d'un candidat avant et après sa mise en examen, choix d'une marque avant et après un message publicitaire, choix des études avant et après la rencontre d'un conseiller d'orientation, analyse des résultats des sondages participatifs,.... Dans tous les cas, on veut savoir si l'action intermédiaire a eu un **effet net** sur les opinions. Par effet net, il faut comprendre que ne sera considéré que l'écart des effectifs des configurations $(A \to B)$ et $(B \to A)$ de la liste ci-dessus : si l'effectif de $(A \to B)$ est supérieur à celui de $(B \to A)$ alors l'action a eu un effet favorable pour B et inversement, si l'effectif de $(A \to B)$ est inférieur à celui de $(B \to A)$ alors l'action a eu un effet favorable pour A au détriment de B.
Fort logiquement les effectifs des individus qui ne changent pas de catégorie ne sont d'aucune utilité pour mesurer un quelconque impact des informations communiquées : pour les personnes classées en $(A \to A)$ et en $(B \to B)$ l'action intermédiaire n'a eu en effet aucune conséquence.

Le fait qu'il s'agisse de la mesure d'un effet net peut se comprendre avec un exemple simple : supposons que les effectifs des deux sous groupes A et B soient égaux et que $(A \rightarrow A)$ et $(B \rightarrow B)$ soient nuls, ce qui signifie que chaque individu a changé d'opinion. Dans ces conditions, nous devons conclure que l'action intermédiaire a été inefficace au regard des positions A et B, et cela alors même qu'individuellement toutes les opinions ont été renversées. Il ne s'agit pas en effet de tester la stabilité individuelle des choix mais simplement l'homogénéité des marges du tableau croisé.

Techniquement, si on représente par un signe '+' les individus passant du sous-groupe A au sous-groupe B, par un signe '-' ceux faisant le chemin opposé et si on ne considère pas ceux restant dans leur sous-groupe initial après l'événement intermédiaire, alors sous l'hypothèse nulle d'absence d'impact de l'action intermédiaire, la probabilité d'observer un changement de sous-groupe dans un sens ou dans l'autre est $p = 1/2$. Soient n^+ et n^- les nombres respectifs de '+' et de '-' et $n_t = n^+ + n^-$, on peut définir une *z-transform* sur la binomiale n^+ :

$$\begin{aligned} z &= \frac{n^+ - n_t p}{\sqrt{n_t p(1-p)}} = \frac{n^+ - 0.5 n_t}{\sqrt{0.25 n_t}} = \frac{n^+ - 0.5(n^+ + n^-)}{\sqrt{0.25 n_t}} \\ &= \frac{0.5(n^+ - n^-)}{0.5\sqrt{n_t}} \\ &= \frac{(n^+ - n^-)}{\sqrt{n_t}} \end{aligned} \tag{2.30}$$

et z tend asymptotiquement vers une gaussienne centrée-réduite ou encore

$$z^2 = \frac{(n^+ - n^-)^2}{n^+ + n^-} \underset{n_t \rightarrow \infty}{\rightarrow} \chi^2(1) \tag{2.31}$$

cette statistique z^2 constitue le chi2 de McNemar et lorsque sa valeur est supérieure à sa valeur critique l'utilisateur peut conclure à un déplacement significatif d'un groupe vers l'autre [40].
Il est implémenté dans la procédure `freq` et est réclamé par l'option `agree` de la commande `table`. Comme pour la statistique de Pearson, il est possible d'obtenir la probabilité critique exacte du χ^2 de McNemar en mobilisant la commande `exact mcnemar` [41]. Cela est évidemment utile lorsque les effectifs sont faibles [42].

Le test de Bowker

Alors que le test de McNemar ne s'applique qu'à des tables 2×2, le test de Bowker le généralise à des tables $k \times k, k \geq 2$. Retenez déjà que pour $k = 2$ les

40. Vous pouvez vérifier aisément que le test de McNemar ne considère effectivement pas la stabilité des choix individuels : supposez qu'initialement 100 personnes soient dans A et 100 dans B. Si à la suite d'une action intermédiaire, les 100 premières rejoignent B et les 100 autres basculent en A, alors la statistique vaut 0.

41. Il s'agit d'une probabilité exacte conditionnelle : la procédure évalue toute les répartitions possibles ayant une probabilité égale ou supérieure à celle obtenue sur l'échantillon de travail pour un effectif donné de changements d'opinion égal à $n^+ + n^-$.

42. Comme toujours, pour que l'approximation asymptotique par la loi du χ^2soit raisonnable il ne faut pas que les effectifs concernés soient trop faibles. Dans la littérature on recommande un effectif total d'au moins 10 individus.

deux tests sont identiques.
Soit une table de fréquences correspondant à des échantillons appariés où chaque individu se situe initialement dans une catégorie i, $i = 1, \ldots, k$, et dans une catégorie j, $j = 1, \ldots, k$ postérieurement à un événement quelconque (résultat d'un test, apport d'une information,...) :

	État final				
État initial	1	2	...	k	Total
1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
k	n_{k1}	n_{k2}	...	n_{kk}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

A cette table contenant des effectifs correspond une table de proportions $p_{ij} = \frac{n_{ij}}{n}$:

	État final				
État initial	1	2	...	k	Total
1	p_{11}	p_{12}	...	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	...	p_{2k}	$p_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
k	p_{k1}	p_{k2}	...	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	...	$p_{.k}$	1

L'hypothèse nulle considérée par le test de Bowker est $H0 : p_{ij} = p_{ji}$ pour tout $i \neq j$, et $i, j = 1, \ldots, k$ versus $H1$: il existe au moins un couple (i, j) avec $i \neq j$ tel que $p_{ij} \neq p_{ji}$. C'est donc un test de symétrie : sous $H0$ les proportions situées au-dessus de la diagonale principale sont égales à celles situées au-dessous. Notez aussi qu'il s'agit d'un test joint.
La statistique de test est donnée par

$$B = \sum_{i}^{k-1} \sum_{j=i+1}^{k} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \tag{2.32}$$

Sous la nulle, la distribution asymptotique de B peut être approximée par celle d'un χ^2 à $k(k-1)/2$ degrés de liberté [43]. En pratique dans la proc `freq`, lorsque l'option `agree` apparaît dans la commande `table`, la statistique de Bowker est évaluée dès lors que la table est plus grande qu'une 2×2 [44].

Après le test lui-même, le point qui mérite d'être discuté est celui de savoir ce que l'on doit faire en cas de rejet : comme tout test joint, on est souvent intéressé à connaître les raisons du rejet, raisons qui ne sont pas explicitées par le test joint. La question est de savoir si on peut identifier les éléments hors-diagonale pour lesquels les proportions se sont déformées entre l'état initial

43. Ici encore cette approximation suppose des effectifs suffisamment importants. Une recommandation est que pour tout i, j on ait $n_{ij} \geq 5$. Notez aussi qu'il n'y a pas de test exact pour la statistique de Bowker contrairement à celle de McNemar.
44. Rappelez-vous que pour une 2×2 les deux statistiques de Bowker et McNemar sont égales.

et l'état final, et ceux pour lesquels il n'y a pas eu de modification ? Une piste peut être suivie pour répondre à cette question : il suffit de revenir à des tests de type McNemar qui vont successivement prendre chacun des éléments hors-diagonale et l'opposer à la totalité des autres éléments.
Un exemple simple doit vous permettre de comprendre la procédure : on dispose de 109 personnes sur lesquelles une caractéristique a été mesurée et répartie en 3 classes, *"inférieure"*, *"normale"*, *"supérieure"*. Après qu'un traitement leur ait été administré la caractéristique a été de nouveau mesurée puis répartie entre les trois mêmes classes. Les données issues de cette expérience vous sont présentées dans la table 2.15. Par exemple, sur les 37 personnes qui avaient une caractéristique jugée *normale* avant le traitement, 16 sont restées avec cette appréciation, 13 ont basculé dans la catégorie *above* et 8 dans la catégorie *below*. Par ailleurs, sur les 44 personnes classées comme *normal* après le traitement, 37 étaient préalablement dans cette même classe, 8 étaient repérées comme *above* et 20 comme *below*.
Remarquez que seules 3 cases situées au-dessus de la diagonale et autant situées au-dessous sont concernées par le test de symétrie. En les examinant, il apparaît à première vue que cette expérience est plutôt défavorable à l'hypothèse de symétrie puisque chacune de ces cases situées sous la diagonale, *i.e.* avant le traitement, a un effectif inférieur à celui contenu dans la case qui lui correspond au-dessus de la diagonale, *i.e.* après le traitement.
L'exécution de la commande `table before*after / agree;` fournit la table 2.16 et un résultat du test de Bowker qui confirme cette impression en permettant de rejeter au seuil usuel de 5% la symétrie de la matrice.

Comme on le précisait au début de cet exemple nous allons maintenant chercher à repérer les éléments qui pourraient être la cause de ce rejet. Pour cela, les trois catégories vont devoir être successivement considérées en créant des configurations qui pourront être étudiées via un test de McNemar. Ces trois configurations sont les suivantes :

1. below *versus* non-below : La table à prendre en compte ici est alors

	État final		
État initial	below	non-below	Total
below	14	32	46
non-below	14	49	63
Total	28	81	109

2. normal *versus* non-normal, avec la table 2×2 suivante

	État final		
État initial	normal	non-normal	Total
normal	16	21	37
non-normal	28	44	72
Total	44	65	109

3. above *versus* non-above au moyen des données

	État final		
État initial	above	non-above	Total
above	12	14	26
non-above	25	58	83
Total	37	72	109

Table of before by after				
	after			
before	below	normal	above	Total
below	14 12.84 30.43 50.00	20 18.35 43.48 45.45	12 11.01 26.09 32.43	46 42.20
normal	8 7.34 21.62 28.57	16 14.68 43.24 36.36	13 11.93 35.14 35.14	37 33.94
above	6 5.50 23.08 21.43	8 7.34 30.77 18.18	12 11.01 46.15 32.43	26 23.85
Total	28 25.69	44 40.37	37 33.94	109 100.00

TABLE 2.15 – Évolution des effectifs entre 3 catégories après un traitement

Symmetry Test		
Chi-Square	DF	Pr > ChiSq
8.3333	3	0.0396

TABLE 2.16 – Résultat du test de Bowker

Les résultats des tests de McNemar réalisées sur ces données sont présentés dans la table 2.17. Trois tests ayant été réalisés, si on désire conserver un seuil de risque de 5%, on sait qu'il faut procéder à un ajustement. La solution la plus simple est de procéder à la correction de Bonferroni pour travailler au seuil de 0.05/3, soit 1.67%. On relève alors un seul mouvement significatif, celui des personnes classées initialement dans la catégorie *below* qui se sont déplacées vers les catégories *normal* et *above* à la suite du traitement.

below *versus* non-below	McNemar's Test		
	Chi-Square	DF	Pr > ChiSq
	7.0435	1	0.0080
normal *versus* non-normal	McNemar's Test		
	Chi-Square	DF	Pr > ChiSq
	1.0000	1	0.3173
above *versus* non-above	McNemar's Test		
	Chi-Square	DF	Pr > ChiSq
	3.1026	1	0.0782

TABLE 2.17 – Résultats des tests de McNemar

2.2.6 Exemple 3 : de l'utilité d'une double correction des copies

Nous disposons des notes données par deux correcteurs différents sur des copies de bac de français [45]. On se demande si l'effectif d'élèves qui ont une note supérieure à la moyenne est dépendant du correcteur.
L'étape data dans les codes qui suivent créé le fichier des notes, et permet la construction de la table (2×2) absolument nécessaire pour conduire le test de McNemar. Les variables `note2` et `note3` contiennent les notes attribuées par chaque correcteur aux 30 copies. Les variables `sup2` et `sup3` valent 1 lorsque la note de la copie est supérieure à 10, et 0 sinon. L'appel de la procédure `freq` avec sa commande `table` renvoie la table 2.18. On y lit que les deux correcteurs sont en accord sur 22 copies (9 copies ont la moyenne chez les deux correcteurs, 13 copies ne l'ont pas). En revanche le troisième correcteur donne la moyenne à 8 copies qui ont une note inférieure à 10 pour le second examinateur. Par ailleurs, les copies n'ayant pas la moyenne chez ce second ne l'ont pas non plus chez le troisième. Dans cet exemple, on aurait donc : $n^+ = 8, n^- = 0$ et donc, en reprenant la formulation de l'équation (2.31), la statistique de McNemar doit valoir : $z^2 = \frac{(8-0)^2}{8+0} = 8$ et est, sous l'hypothèse nulle d'absence d'écart net dans les nombres de copies ayant ou n'ayant pas la moyenne, la réalisation d'un χ^2 à un degré de liberté. C'est ce que confirme la lecture de la table 2.19 qui montre également qu'aux seuils de risques usuels la probabilité critique tant asymptotique qu'exacte conduit au rejet de la nulle : les effectifs de reçus au bac de français dépendent du correcteur.

```
data notes;
  input note2 note3 @@;
  if note2>=10 then sup2=1; else sup2=0;
  if note3>=10 then sup3=1; else sup3=0;
cards;
12 13
9 10
8 9
4 10
11 11
5 9
6 13
9 14
6 13
8 12
12 13
9 8
6 7
9 7
4 6
8 9
9 6
```

45. Les données sont tirées d'une étude intitulée "A quelques points près : les notes au baccalauréat" réalisée par Annie Uhry et Claudine Schwartz, travail accessible à l'adresse http ://www.statistix.fr/IMG/pdf/Variabilite_des_notes.pdf
Dans cette étude, on peut trouver les notes de 10 correcteurs différents attribuées à 30 copies identiques. Nous avons pris au hasard celles de leurs 2ème et 3ème correcteurs.

Frequency Percent Row Pct Col Pct	Table of sup2 by sup3		
	sup3		
sup2	0	1	Total
0	13 43.33 61.90 100.00	8 26.67 38.10 47.06	21 70.00
1	0 0.00 0.00 0.00	9 30.00 100.00 52.94	9 30.00
Total	13 43.33	17 56.67	30 100.00

TABLE 2.18 – Copies ayant ou n'ayant pas la moyenne - Croisement des avis de deux correcteurs

McNemar's Test	
Statistic (S)	8.0000
DF	1
Asymptotic Pr > S	0.0047
Exact Pr >= S	0.0078

TABLE 2.19 – Statistique de McNemar - probabilités critiques asymptotique et exacte

```
5 7
7 10
14 14
16 12
12 11
14 12
12 14
6 5
9 8
8 7
6 10
9 8
10 11
;run;
proc freq data=notes;
  table sup2*sup3 / agree;
  exact mcnemar;
run;
```

2.2.7 Le test binomial d'une proportion

Le test de signe vu ci-dessus a été utilisé pour tester une valeur particulière pour la médiane d'un ensemble d'observations. Lorsque l'interrogation porte sur une proportion autre que 1/2, on peut passer par la procédure `freq` et l'op-

tion `binomial` de la commande `table` [46].
Le cas type est celui où les n observations d'une variable ont été ont été réparties entre plusieurs classes, $c_1, c_2, \ldots, c_k$ d'effectifs respectifs $n_1, n_2, \ldots, n_k$ vérifiant $\sum_{i=1}^{k} n_i = n$, et où le vecteur des probabilités d'appartenance aux diverses classes est $p_1 p_2, \ldots, p_k$ avec $\sum_{i=1}^{k} p_i = 1$. L'hypothèse nulle est que la probabilité d'appartenance à une classe est égale à une certaine valeur p_0, soit par exemple : $H0 : p_i = p_0$ où i est un entier tel que $1 \leq i \leq k$.

L'option `binomial` requiert que l'on précise plusieurs éléments :

- La classe concernée parmi les k possibles. Le plus simple et le moins dangereux est d'appliquer un format sur la variable pour imposer l'intitulé des k classes. Il suffit ensuite de préciser `level='intitulé de la classe à considérer'`. En l'absence de format, on peut aussi spécifier `level=`k_0, où k_0 est un entier positif indiquant le rang de création de la classe que l'on désire retenir, sachant que l'ordre de création dépend du choix fait sur l'option `order` de la ligne d'appel ainsi que nous l'avons déjà vu [47]. Par défaut, si `level=` n'est pas utilisé, la procédure va retenir la première classe créée [48].
- La valeur de la probabilité supposée p_0, via `P=` . En l'absence de cette précision, la procédure force $p_0 = 0.5$.
- Le seuil de confiance à employer pour construire des intervalles de confiance sur la proportion estimée, $\hat{p}_i$, avec `ALPHA=` . Par défaut cette valeur est de 95%.
- le type d'intervalle de confiance à construire, avec `CL=type`, *type* étant le nom de l'intervalle à construire choisi parmi `CLOPPERPEARSON`, `JEFFREYS`, `LIKELIHOODRATIO`, `LOGIT`, `WALD`, `WILSON` [49].

Soit par exemple : $H0$: 'La probabilité d'appartenir à la i$^{\text{ème}}$ classe intitulée 'azerty' de la variable X est $p\_0 = 0.3$. Les commandes à exécuter pourraient être de la forme

```
proc freq data=...;
table X /
  binomial(
    level='azerty' p=0.3 cl=
    (wald clopperpearson jeffreys
    likelihoodratio logit wilson)
  );
run;
```

46. On peut toutefois signaler que l'appel de `freq` pour faire un test de proportion égale à 0.5 peut différer du résultat d'un test de signe mené avec la procédure `univariate` sur la valeur d'une médiane car cette dernière élimine les observations égales à M_0, alors qu'elles sont conservées dans les calculs dans la proc `freq`.

47. Pour mémoire, nous avons les choix `order=data/formatted/freq/internal`.

48. Qui correspond à la première ligne dans le tableau de sortie.

49. Les curieux pourront trouver dans l'aide de la procédure `freq` sous l'onglet `détails`, rubrique `Statistical Computations`, section `Binomial Proportion` les expressions détaillées de tous ces intervalles sachant que vous devez déjà connaître celle de Wald pour laquelle, par exemple, sur la i$^{\text{ème}}$ classe, $\hat{p}_i = n_i/n$, $s_{\hat{p}_i} = \sqrt{\hat{p}_i(1-\hat{p}_i)/n}$ avec l'intervalle au seuil de confiance α donné par $\hat{p}_i \pm z_{\alpha/2} s_{\hat{p}_i}$. Sachez que l'aide de SAS semble recommander l'emploi de type=WILSON, et que l'utilisateur peut utiliser la syntaxe `type=(liste de noms d'intervalles)` afin de vérifier si le choix de tel ou tel est déterminant ou non.

La statistique affichée est une *z-transform* sous $H0$ de la probabilité estimée [50] [51]

$$Z = \frac{\hat{p}_i - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{2.33}$$

Enfin, il est possible de calculer la probabilité critique exacte ainsi que les bornes exactes des intervalles de confiance sur un test de proportion via la commande

```
exact binomial
```

qui ne doit généralement pas être utilisée sans son option `maxtime=`

50. Il est possible de demander l'application d'une correction pour continuité en précisant `binomial(correct)` ou directement `binomialc`.

51. Notez encore que l'option `VAR=SAMPLE` de `binomial` construirait l'écart-type de la z-transform avec la proportion estimée et non pas la proportion imposée par l'hypothèse nulle, ainsi avec `binomial(p0=0.3 level='azerty' var=sample)` on aurait $Z=(\hat{p}_i - p_0)/\sqrt{\hat{p}_i(1-\hat{p}_i)/n}$. En théorie, c'est bien (2.33) qui doit être utilisée.

Chapitre 3

Tests d'égalité de deux distributions

Le cadre général est le suivant : On dispose de deux échantillons, $X = \{x_1, x_2, \ldots, x_{n_x}\}$ et $Y = \{y_1, y_2, \ldots, y_{n_y}\}$ constitué chacun de réalisations indépendantes de deux aléatoires ayant des fonctions de répartition inconnues F_x et F_y.

Une question fréquente est alors celle de savoir si les deux aléatoires possèdent la même distribution. Le premier test présenté, appelé test de la médiane s'inspire du test du signe déjà vu et ne présente aucune difficulté théorique supplémentaire par rapport à ce dernier. Les deux tests suivants, de Mann-Whitney et de la somme des rangs de Wilcoxon, sont couramment utilisés pour répondre à la question qui nous intéresse ici. Par ailleurs, même si au premier abord leur approche est différente, ils sont en fait équivalents : on n'a donc pas à privilégier l'un plutôt que l'autre.
Par la suite, nous retrouverons les tests EDF déjà exposés lorsqu'il s'agissait de regarder la validité d'une hypothèse concernant la distribution d'un ensemble d'observations en construisant une distance entre les fonctions de répartition théorique et empirique. Maintenant, ils vont être employés pour comparer F_x et F_y à partir de la distance de leurs deux estimations respectives $\hat{F}_x$ et $\hat{F}_y$.
Tous ces premiers tests s'appliquent notamment lorsque l'on dispose des observations individuelles $x_i, i = 1, \ldots, n\_x$ et $y_i, i = 1, \ldots, n_y$. Par la suite, dans le chapitre 4, nous aborderons les tests utilisant les données des ensembles X et Y regroupées en classes et retrouverons notamment le test de χ^2 de Pearson, mais aussi le test de Fisher exact tout particulièrement utile lorsque les effectifs des classes font douter de l'adéquation de la distribution asymptotique de la statistique de Pearson.

3.1 Le test de la médiane

On dispose de deux échantillons X et Y de tailles respectives n_x et n_y, avec une hypothèse nulle $Ho : F_x = F_y$ versus une alternative soit bidirectionnelle $H1 : F_x \neq F_y$, soit unidirectionnelle $H1 : F_x > F_y$ ou $H1 : F_x < F_y$.
Pour réaliser ce test, on créé des scores, $s(r_i)$ à partir des rangs des observations

dans l'échantillon mélangé. Ces scores sont définis par :

$$s(r_i) = \begin{cases} 1 & \text{si} \quad r_i > (n+1)/2 \\ 0 & \text{si} \quad r_i \leq (n+1)/2 \end{cases} \tag{3.1}$$

La statistique de test, S, est la somme des scores de l'échantillon de référence [1]. Il s'agit donc simplement du nombre d'observations de cet échantillon qui sont supérieures à la médiane. Si X est cet échantillon de référence, on a

$$S = \sum_{i=1}^{n} z_i s(r_i), \text{avec } z_i = 1 \text{ si } x_i \in X \text{ et } z_i = 0 \text{ sinon} \tag{3.2}$$

$$E[S] = \frac{n_x}{n} \sum_{i=1}^{n} s(r_i), \text{ et} \tag{3.3}$$

$$Var[S] = \frac{n_x n_y}{n(n-1)} \sum_{i=1}^{n} (s(r_i) - \bar{s})^2 \tag{3.4}$$

ou $n = n_x + n_y$ et $\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(r_i)$ est le score moyen. Pour le calcul de la probabilité critique asymptotique, la procédure recourt alors à une *z-transform* :

$$z = \frac{s - E[S]}{\sqrt{Var[S]}} \tag{3.5}$$

Si $\tilde{z}$ est une gaussienne centrée réduite, la probabilité critique asymptotique bidirectionnelle correspond à $Pr(|\tilde{z}| > z)$, et l'unidirectionnelle à $Pr(\tilde{z} > z)$ si $z > 0$ ou à $Pr(\tilde{z} < z)$ si $z < 0$. Ces probabilités sont affichées dans l'output de la procédure, ainsi d'ailleurs que sa transformée z^2 qui est évidemment sous $H0$ la réalisation d'une statistique de khi-deux à 1 degré de liberté.

La statistique S possédant une distribution hypergéométrique, il est possible d'obtenir sa probabilité critique exacte via la commande `exact median / maxtime=` . On sait aussi que lorsque les temps de calculs sont trop longs, une estimation par bootstrap peut être réclamée, avec la syntaxe usuelle, par exemple : `exact median / seed=123;`

3.2 Les tests de Mann-Whitney et de la somme des rangs de Wilcoxon

3.2.1 Le test de Mann-Whitney

La logique qui sous-tend ce test est très simple : supposons que l'on mélange les deux ensembles d'observations, et que l'on classe par ordre croissant les valeurs de l'échantillon ainsi créé alors, si $H0$ est vraie, nous devrions observer une alternance régulière tout au long du support de l'échantillon joint des valeurs prises dans X d'une part et dans Y d'autre part. En revanche, l'apparition de zones de concentration d'observations issues de X ou de Y est défavorable à l'hypothèse nulle.

1. Par défaut le plus petit des deux.

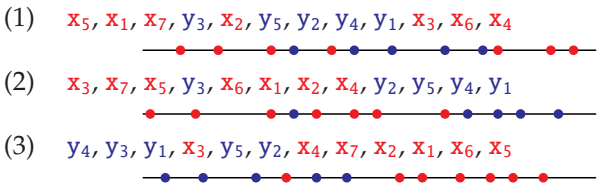


FIGURE 3.1 – Test de Mann-Whitney - Exemples de mélanges d'échantillons

On illustre le raisonnement avec les exemples suivants : soit $n_x = 7$ et n_y=5. Considérons les trois configurations présentées dans la figure 3.1 pour les valeurs classées de l'échantillon joint. Elles sont toutes défavorables à l'hypothèse nulle :

(1) Dans le premier cas, les valeurs prises dans X occupent plutôt les rangs extrêmes dans le mélange. On peut soupçonner que la distribution de X est plus étalée que celle de Y : elles auraient un même centre, mais la variance de celle-ci serait plus petite que la variance de celle-là.

(2) Dans ce mélange, les rangs des observations de X sont plutôt faibles alors que les observations tirées de Y occupent les rangs élevés. La densité f_y pourrait être décalée à droite par rapport à f_x.

(3) C'est l'opposé du précédent les observations prises dans X sont en majorité sur les rangs élevés dans le mélange : f_y serait décalée à gauche par rapport à f_x.

Le test de Mann-Whitney est simplement le nombre de fois où un y précède un x dans l'échantillon mélangé et classé, soit :

$$U = \sum_{i=1}^{n_x} \sum_{i=1}^{n_y} \mathbb{I}(x_i > y_i) \tag{3.6}$$

Dans les trois exemples précédents, nous avons ainsi

(1) U=0+0+0+1+5+5+5=16

(2) U=0+0+0+1+1+1+1=4

(3) U=3+5+5+5+5+5+5=33

Comme on peut s'en douter, la statistique U est à même de bien révéler les cas correspondant à un décalage d'une densité par rapport à une autre, en prenant soit des valeurs faibles, soit des valeurs élevées, comme dans les configurations (2) et (3) précédentes. Par contre, lorsque les centres des deux distributions seront proches, les différences de répartition seront moins aisément repérables en raison des compensations qui vont se produire dans le calcul de U avec les x de rangs faibles, qui conduiraient à une valeur de la statistique faible, et les x de rangs élevés, qui induiraient plutôt une grande valeur de U.
Si on note W_x (resp. W_y) la somme des rangs des observations de X (resp. de Y) dans l'échantillon mélangé, on peut montrer que la statistique U de (3.6) est

encore donnée par

$$U = \min(U_x, U_y), \text{ avec} \tag{3.7}$$

$$U_x = n_x n_y + \frac{n_x(n_x+1)}{2} - W_x, \text{ et} \tag{3.8}$$

$$U_y = n_x n_y + \frac{n_y(n_y+1)}{2} - W_y \tag{3.9}$$

3.2.2 Le test de Wilcoxon

Le point de départ est le même que celui du test de Mann-Whitney, à savoir un mélange des deux échantillons X et Y. Ensuite, on calcule la somme des rangs des observations constituant X, somme qui constitue la statistique de test W. Comme on peut s'en douter, ce test va aussi hériter du défaut de U : il est mieux adapté à la détection des situations où $F_y(y) = F_x(y+\theta)$ ou $F_y(y) = F_x(y-\theta)$, avec $\theta > 0$, qu'à celles où $F_y(y) \neq F_x(y)$ mais pour lesquelles les centres de distributions seront proches.
La distribution de W est connue sous $H0 : F_x(x) = F_y(x)$: l'échantillon joint est de taille $n_x + n_y$ et il y a donc $(n_x + n_y)!$ ordres possibles. Le nombre de façons de placer les $x_i, i = 1, \ldots, n_x$ est

$$\binom{n_x+n_y}{n_x} = \frac{(n_x+n_y)!}{n_x! n_y!}$$

tous ces rangements étant équiprobables. Enfin, si on note w la valeur de la statistique W obtenue sur l'échantillon joint, et si k_W est, parmi tous les rangements possibles le nombre de ceux pour lesquels la somme des rangs des observations des x_i vaut w, alors :

$$Pr[W = w] = \frac{k_w}{\binom{n_x+n_y}{n_x}} \tag{3.10}$$

Ce calcul exact n'est pas trop couteux pour de faibles valeurs de n_x et n_y. Pour des tailles importantes, on peut basculer sur une approximation de type *z-transform* : si on note Z_i une indicatrice valant 1 si une observation x_i est en i$^{\text{ème}}$ position et 0 sinon, il vient :

$$\begin{aligned} E[W] &= E\left(\sum_{i=1}^{n_x+n_y} iZ_i\right) \\ &= E(Z_i) \sum_{i=1}^{n_x+n_y} i \\ &= \frac{n_x}{n_x+n_y} \frac{(n_x+n_y)(n_x+n_y+1)}{2} \\ &= \frac{n_x(n_x+n_y+1)}{2} \end{aligned} \tag{3.11}$$

Un calcul similaire montre que :

$$\text{Var}(W) = \frac{n_x n_y (n_x+n_y+1)}{12} \tag{3.12}$$

Au final, on considère la valeur de

$$z = \frac{W - \frac{n_x(n_x+n_y+1)}{2}}{\sqrt{\frac{n_x n_y(n_x+n_y+1)}{12}}} \tag{3.13}$$

qui, sous $H0$, est la réalisation d'une gaussienne centrée réduite [2].

Enfin, un calcul simple montre que $U = W - n_x(n_x + 1)/2$, ce qui démontre l'équivalence des deux statistiques.Dans le cas (1) considéré ci-dessus, il vient par exemple, $W = 1 + 2 + 3 + 5 + 10 + 11 + 12 = 44$ et $44 - 7 \times 8/2 = 16$, ce qui correspond bien à la valeur de la statistique de Mann-Whitney trouvée alors.

L'obtention de la statistique W de Wilcoxon s'effectue dans la procédure `NPAR1WAY`. Il suffit de mettre le mot clef `wilcoxon` dans la ligne d'appel de celle-ci.

généralisations des test U et W

Les tests de Man-Whitney et de Wilcoxon appartiennent à une classe de tests non paramétriques fondés sur les rangs. Le principe général de construction est de construire les rangs des observations, r_i, dans l'échantillon mélangé puis de calculer une statistique dont l'expression générale est, par exemple sur deux échantillons X et Y, de la forme

$$S = \sum_{i=1}^{n_x+n_y} c_i s_i \tag{3.14}$$

où s_i est un score que l'on attribue à la i$^{\text{ème}}$ observation, score obtenu à partir des rangs et c_i une indicatrice de l'appartenance à l'un ou l'autre des échantillons initiaux. Or, si on connaît la distribution commune à X et Y sous $H0$, il est possible de choisir des scores optimaux pour cette distribution. Par exemple, la condition $s_i = r_i$, qui donne la statistique de Wilcoxon, est optimale pour une distribution logistique. Si la distribution est une exponentielle double, alors les scores optimaux sont des scores médians définis par $s_i = 1$ si $r_i > (n_x + n_y)/2$ et $s_i = 0$ sinon. Adaptés aux distributions exponentielles ou aux distributions des valeurs extrêmes, les scores de Savage, sont donnés par $s_i = \sum_{i=1}^{r_i} \left(\frac{1}{n_x+n_y-i+1}\right) - 1$. Les scores de Van der Waerden sont optimaux lorsque la distribution est gaussienne, ils sont définis par $s_i = \Phi^{-1}\left(\frac{r_i}{n+1}\right)$. Ces généralisations sont disponibles dans la procédure `NPAR1WAY` avec les options `wilcoxon`, `median`, `savage` , et il est aussi possible de calculer leur probabilité critique exacte, via, comme d'habitude, la commande `exact` *`nom de la statistique`* [3].

l'estimation de Monte Carlo de la probabilité critique exacte

Comme on le sait, les temps de calculs des probabilités critiques exactes peuvent devenir déraisonnable. Dans ce cas, si on suppose que les propriétés

2. Par défaut, `npar1way` va intégrer une correction pour la continuité. Il est possible de ne pas appliquer cette correction en mettant l'option `correct=no` dans la ligne d'appel de la procédure.

3. La liste des statistiques concernées peut être consultée sous l'onglet `syntax` de la procédure `NPAR1WAY`, section `exact statement`, section `Statistics Options`.

asymptotiques ne sont pas satisfaites alors la procédure permet la mise en oeuvre de techniques de Monte Carlo. Il suffit de mettre l'option `MC` dans la commande `exact`. Implicitement, les options `N=` , `ALPHA=` et `SEED=` activent l'option `mc`. On rappelle qu'elles gèrent respectivement le nombre d'échantillons simulés par bootstrap, le seuil de confiance utilisé pour construire un intervalle de confiance sur la probabilité critique, le contrôle des tirages des échantillons bootstrap autorisant la réplication des résultats.

3.3 Les tests EDF

Ainsi que nous l'avons vu dans la présentation des tests sur un seul échantillon, les tests EDF font appel aux fonctions de répartition empiriques. Les observations constituant les échantillons doivent être mesurées au moins sur une échelle ordinale.

3.3.1 le test de Kolmogorov-Smirnov

On dispose de k échantillons indépendants eux-mêmes constitués de réalisations indépendantes d'une aléatoire propre à chacun, soit :

$$\begin{aligned} X_1 &= \{x_{11}, x_{12}, \ldots, x_{1n_1}\} \\ X_2 &= \{x_{21}, x_{22}, \ldots, x_{2n_2}\} \\ &\vdots \\ X_k &= \{x_{k1}, x_{k1}, \ldots, x_{kn_k}\} \end{aligned}$$

La procédure va construire les répartitions empiriques pour chacun des échantillons et celle de l'échantillon mélangé $X = X_1 \cup X_2 \cup \ldots \cup X_k$ selon

$$\hat{F}_{X_i}(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{I}(x_{ij} \leq x), i = 1, \ldots, k, \text{ et} \tag{3.15}$$

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbb{I}(x_{ij} \leq x) = \frac{1}{n} \sum_{i=1}^{k} n_i \hat{F}_{X_i}(x) \tag{3.16}$$

avec $n = n_1 + n_2 + \ldots + n_k$. La statistique de Kolmogorov-Smirnov est égale à l'écart maximal observé entre les répartitions empiriques des k classes et celle de l'échantillon mélangé [4] :

$$KS = \max_i \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \left(\hat{F}_{X_i}(x_{.j}) - \hat{F}(x_{.j}) \right)} \tag{3.17}$$

où $x_{.j}, j = 1, 2, \ldots, n$ sont les observations de l'échantillon mélangé. La valeur de la statistique affichée est $KS_a = KS\sqrt{n}$.

4. On comprend alors l'intérêt de la référence à cet échantillon mélangé : cela permet de tester des differences de fonctions de répartition sur plus de deux échantillons.

Lorsqu'il n'y a que deux classes, X et Y, la procédure évalue la statistique de Kolmogorov-Smirnov comme :

$$D = \max |\hat{F}_x(x_{.j}) - \hat{F}_y(x_{.j})|, j = 1, 2, \ldots, n_x + n_y \tag{3.18}$$

La probabilité critique asymptotique de la statistique est évaluée grâce au théorême de Kolmogorov au moyen de l'équation (2.3). Elle approxime la probabilité d'obtenir une valeur de D supérieure à celle observée sachant que $H0$:'les fonctions de répartition des sous-groupes sont égales' est vraie. Comme d'habitude, on rejette donc l'hypothèse nulle si cette probabilité est inférieure au seuil de risque retenu par l'utilisateur. Il s'agit d'une probabilité bilatérale opposant $H0 : F_x(z) = F_y(z)$ à $H1 : F_x(z) \neq F_y(z)$.
Retenez qu'avec deux classes, il est possible de réaliser un test unilatéral, *i.e.* d'opposer $H0 : F_x(z) = F_y(z)$ à $H1 : F_x(z) < F_y(z)$ ou à $H1 : F_x(z) > F_y(z)$, et la procédure `NPAR1WAY` calcule alors les deux statistiques

$$D^+ = \max\left(\hat{F}_x(x_{.j}) - \hat{F}_y(x_{.j})\right), j = 1, 2, \ldots, n_x + n_y \text{ et} \tag{3.19}$$

$$D^- = \max\left(\hat{F}_y(x_{.j}) - \hat{F}_x(x_{.j})\right), j = 1, 2, \ldots, n_x + n_y \tag{3.20}$$

avec, pour toutes ces statistiques de type D, affichage des probabilités critiques asymptotiques et, si réclamées par la commande `exact`, celui de leur probabilités critiques exactes.

3.3.2 le test de Cramer-von Mises

Nous connaissons déjà ce test : la distance entre deux fonctions n'est pas appréciée comme avec KS par un écart maximal, mais par la somme des carrés des différences des écarts calculés sur la totalité des observations. En reprenant les notations de la sous-section précédente, la statistique est donnée par :

$$CM = \frac{1}{n^2} \sum_{i=1}^{k} \left(n_i \sum_{j=1}^{p} t_j \left(\hat{F}_i(x_{.j}) - \hat{F}(x_{.j}) \right)^2 \right) \tag{3.21}$$

où p est le nombre d'observations distinctes, et t_j les effectifs d'observations égales pour chaque valeur distincte. Les écarts sont ceux de la fonction de répartition empirique de chacune des classes à celle de l'échantillon mélangé. La statistique peut donc être évaluée même en présence de plus de deux classes.
En raison de la prise des carrés des écarts, ce test est obligatoirement bilatéral.

En plus de CM la procédure donne une statistique asymptotique : $CM_a = n \times CM$ pour laquelle on peut trouver des tables donnant les probabilités critiques.

3.3.3 le test de Kuiper

Alors que les tests EDF précédents peuvent être appliqués sur plus de deux échantillons, le test de Kuiper est réalisé uniquement lorsqu'on compare deux échantillons. La statistique est donnée par :

$$K = \max_z \left[F_x(z) - F_y(z) \right] - \min_z \left[F_x(z) - F_y(z) \right] \tag{3.22}$$

Sa valeur asymptotique, sur laquelle on peut estimer une probabilité critique est

$$K_a = K\sqrt{n_x n_y} n_x + n_y \tag{3.23}$$

Cette probabilité critique asymptotique est celle d'observer une valeur supérieure à K_a sous l'hypothèse $H0 : F_x = F_y$. On rejette donc $H0$ si elle est inférieure au seuil de risque retenu.

3.3.4 Exemple 1 : notes à la session 1 des examens du M1 ESA selon la provenance des étudiants

Les inscriptions en première année du Master ESA ouvert dans le département d'économie de l'Université d'Orléans sont de droit pour les titulaires de la licence d'économétrie également offerte au sein de ce département. Des candidatures en provenance d'autres Universités sont également autorisées sur lesquelles une commission de validation des acquis donne un avis. Une des conséquences est que le pourcentage de mentions "Assez Bien" ou supérieures obtenues à la licence est plus élevé pour les M1 ESA provenant de l'extérieur que pour ceux issus de la licence d'économétrie, pour lesquels une mention "Passable" suffit donc à l'admission. Ceci peut laisser présager que les résultats en M1 seront meilleurs pour les extérieurs. En revanche, les titulaires de la licence d'économétrie d'Orléans ont une meilleure connaissance de l'environnement de travail, peuvent être plus proches de camarades ayant déjà intégrés le Master, ou/et ont suivi des cours de licence qui anticipent le contenu des cours de Master peut être mieux que ne le font les enseignements dispensés dans d'autres Universités. Ces éléments plaident plutôt pour un avantage des locaux qui pourrait se traduire par de meilleurs résultats en M1. Pour tenter de trancher entre ces énoncés contradictoires, nous avons relevé les notes en première session des examens de M1 de l'année universitaire 2015-2016. Au total, 38 notes sont conservées avec lesquelles deux groupes sont constitués : les titulaires de la licence d'économétrie d'Orléans d'une part (18 étudiants), les extérieurs d'autre part [5](20 étudiants). La création de la table de travail est réalisée au moyen d'une étape `data` dont le contenu est donné ci-après. Outre la variable `note`, nous avons également une variable `origine` dont les modalités, "Extérieur" et "L3_Orléans" vont définir les sous-échantillons au moyen d'une instruction `class` présente dans l'appel de la procédure `NPAR1WAY` dans laquelle sont implémentés tous les tests qui viennent d'être présentés.

```
data notesM1;
input origine $1-10 note;
cards;
Extérieur  11.35
Extérieur  10.009
L3_Orléans 6.481
L3_Orléans 9.316
L3_Orléans 10.953
Extérieur  13.859
L3_Orléans 8.415
```

5. Tous les étudiants en provenance d'autres filières, comme par exemple les MASE de l'Université d'Orléans ont été exclus, de même que les défaillants aux examens et les redoublants.

```
Extérieur  10.597
Extérieur  8.747
L3_Orléans 15.066
L3_Orléans 11.126
L3_Orléans 10.732
L3_Orléans 11.979
Extérieur  9.538
L3_Orléans 10.598
L3_Orléans 12.435
Extérieur  14.703
Extérieur  10.801
L3_Orléans 11.465
L3_Orléans 12.022
L3_Orléans 8.672
Extérieur  11.324
Extérieur  7.682
L3_Orléans 13.401
L3_Orléans 9.548
Extérieur  13.543
Extérieur  9.659
L3_Orléans 10.514
L3_Orléans 15.401
Extérieur  14.988
L3_Orléans 8.101
Extérieur  11.768
Extérieur  15.508
Extérieur  10.937
Extérieur  8.092
Extérieur  9.636
Extérieur  13.315
Extérieur  8.668
;
run;
proc npar1way data=notesM1;
class origine;
var note;
exact / maxtime=60;
run;
```

La syntaxe d'appel de la proc `npar1way` ne pose pas de difficultés : une commande `class` qui donne le nom de la variable unique dont les modalités, qui peuvent être plus de deux, vont définir des sous-échantillons sur les observations de ou des variables numériques référencées dans la commande `var`. Si besoin, une commande `freq` *variable*, où *variable* est le nom d'une variable numérique contenant les effectifs des observations présentes dans la table utilisée en entrée peut apparaître. Et enfin une commande `exact` pour le calcul des probabilités critiques exactes de plusieurs statistiques [6], avec ses options, identiques à celles vues avec la proc `univariate`, à savoir `maxtime=` et `mc`, `seed=`,

6. Parmi lesquelles Wilcoxon et Kolmogorov-Smirnov

Analysis of Variance for Variable note Classified by Variable origine		
origine	N	Mean
Extérieur	20	11.236200
L3_Orléans	18	10.901389

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	1	1.061986	1.061986	0.1924	0.6636
Within	36	198.732503	5.520347		

TABLE 3.1 – Moyenne des notes en session 1 du M1 ESA - Étudiants provenant de la licence d'économétrie locale ou de l'extérieur

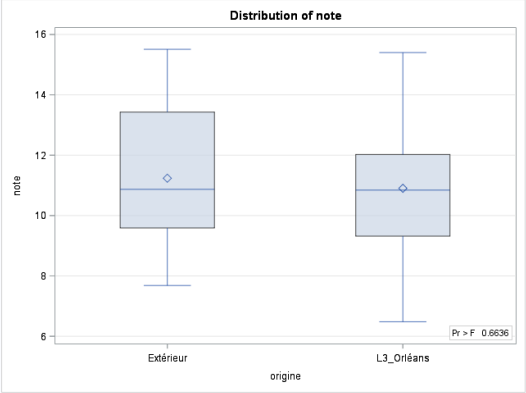


FIGURE 3.2 – Boîte à moustache - Distribution des notes en session 1 du M1 ESA - Étudiants provenant de la licence d'économétrie locale ou de l'extérieur

`alpha=, n= .`

Dans cet exemple, comme on ne précise pas dans la commande `exact` le nom d'une statistique particulière, la procédure va faire les calculs pour toutes celles autorisant l'évaluation de leur probabilité critique exacte [7]. On récupère entre autres les informations de la table 3.1 faisant apparaître une moyenne légèrement supérieure pour les étudiants provenant d'autres Universités que celle d'Orléans. Cependant la table d'analyse de la variance également affichée ne permet pas, avec un fisher égal à 0.1924 de rejeter l'hypothèse d'égalité des espérances des deux groupes. Des graphiques de type "boîte à moustache" sont également tracés pour chacun des sous-échantillons d'intérêt [8]. Si on en juge par l'écart interquartile on peut observer une plus grande concentration de la distribution des notes des étudiants orléanais qui aurait cependant une plus grande étendue.

En ce qui concerne les tests non paramétriques, le premier présenté est celui de Wilcoxon. Dans le tableau 3.2 on va trouver la valeur des sommes des rangs

7. Comme toujours, pensez alors à préciser un nombre de secondes via l'option `maxtime= .`

8. Pour rappel : les côtés bas et haut du rectangle correspondent respectivement au premier et troisième quartile de la distribution étudiée, la hauteur est donc égale à l'écart interquartile Q3-Q1, le trait horizontal à l'intérieur du rectangle indique l'emplacement de la médiane, le symbole du losange celui de la moyenne, enfin, la "moustache" donne l'étendue de la distribution ,*i.e.* les positions de ses valeurs minimale et maximale.

Wilcoxon Scores (Rank Sums) for Variable note Classified by Variable origine					
origine	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Extérieur	20	398.0	390.0	34.205263	19.900000
L3_Orléans	18	343.0	351.0	34.205263	19.055556

TABLE 3.2 – Somme des rangs signés de Wilcoxon - Étudiants provenant de la licence d'économétrie locale ou de l'extérieur

signés observée et attendue sous l'hypothèse nulle dans les sous-échantillons. On peut voir que ces deux sommes sont proches, signalant l'absence d'hétérogénéité des rangs des observations dans l'échantillon mélangé en fonction de leur provenance. C'est ce que confirment les résultats des tests eux-mêmes présentés dans la table 3.3 : tant la probabilité critique asymptotique de la réalisation de la *z-transform*, tirée d'une distribution gaussienne, que celle calculée à partir d'une distribution de student [9], et que la probabilité critique exacte, ne permettent pas de rejeter l'hypothèse nulle d'égalité des deux distributions de notes [10].
Les résultats du test de la médiane sont ensuite présentés. Dans la table 3.4, on va trouver la somme des scores, ici le nombre d'observations supérieures à la médiane de l'échantillon joint dans chaque échantillon, et la somme espérée sous H0. Dans cet exemple, l'égalité parfaite des deux implique une valeur nulle pour la statistique z et aussi naturellement pour le χ^2 qui en découle. Dans ces conditions, ce test ne rejette évidemment pas l'hypothèse nulle d'égalité des deux distributions.
En ce qui concerne

Pour le test de Kolmogorov-Smirnov, une première table (3.5) présente les écarts des fonctions de répartition empiriques de chaque échantillon à leur moyenne. Une autre donne les valeurs des statistiques KS, définie en (3.17), et KS_a [11] ainsi que les statistiques D, D^+ et D^- avec que leurs probabilités critiques [12] permettant de faire un test de l'hypothèse nulle $H0$ d'égalité des deux fonctions de répartition versus des alternatives $H1$ successivement bilatérale et unilatérales. Ces statistiques présentées dans la table 3.6, ne permettent pas avec nos données de rejeter $H0$ aux seuils de risque usuels.

On notera que les 60 secondes allouées au temps de calcul des probabilités critiques exactes n'ont pas été suffisantes, ce qui explique l'indication d'un signe de valeur manquante dans la table pour ce qui concerne leurs valeurs [13]

9. On se rappelle que dans la z-transform, on divise une statistique centrée sous $H0$ par une estimation de son écart-type. Cette opération est responsable d'une modification de sa distribution qui, de gaussienne, devient une student.
10. On obtient la même conclusion en regardant les résultats, non reproduits ici, des tests généralisant Wilcoxon, à partir des scores de Van der Waerden et de Savage qui sont également donnés par défaut par la procédure `npar1way`.
11. Pour information, on a obtenu $KS = 0.0693$ et $KS_a = 0.4275$.
12. Voir l'équation (3.18), et les deux équations qui la suivent
13. On peut noter que la commande `maxtime=s` où s est un nombre de secondes, ne définit pas un temps maximum global de calcul pour les probabilités critiques exactes : la procédure alloue ce temps s à chacune des statistiques pour laquelle ces probabilités sont demandées. Rappelez-vous qu'il peut être avantageux de remplacer la ligne de commande `exact` du texte par `exact / mc` de sorte à obtenir des estimateurs des probabilités critiques par bootstrap.

Wilcoxon Two-Sample Test	
Statistic (S)	343.0000
Normal Approximation	
Z	-0.2193
One-Sided Pr < Z	0.4132
Two-Sided Pr > |Z|	0.8264
t Approximation	
One-Sided Pr < Z	0.4138
Two-Sided Pr > |Z|	0.8276
Exact Test	
One-Sided Pr <= S	0.4142
Two-Sided Pr >= |S - Mean|	0.8285
Z includes a continuity correction of 0.5.	

TABLE 3.3 – Test de Wilcoxon - Distribution des notes en session 1 du M1 ESA selon la provenance des étudiants

Median Scores (Number of Points Above Median) for Variable note Classified by Variable origine

origine	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Extérieur	20	10.0	10.0	1.559626	0.50
L3_Orléans	18	9.0	9.0	1.559626	0.50

Median Two-Sample Test	
Statistic	9.0000
Z	0.0000
One-Sided Pr < Z	0.5000
Two-Sided Pr > |Z|	1.0000

Median One-Way Analysis	
Chi-Square	0.0000
DF	1
Pr > Chi-Square	1.0000

TABLE 3.4 – Test de la médiane - Étudiants provenant de la licence d'économétrie locale ou de l'extérieur

Kolmogorov-Smirnov Test for Variable note Classified by Variable origine

origine	N	EDF at Maximum	Deviation from Mean at Maximum
Extérieur	20	0.750000	-0.294219
L3_Orléans	18	0.888889	0.310135
Total	38	0.815789	
Maximum Deviation Occurred at Observation 24			
Value of note at Maximum = 13.4010			

TABLE 3.5 – Test de Kolmogorov-Smirnov - Distribution des notes en session 1 du M1 ESA selon la provenance des étudiants

Kolmogorov-Smirnov Two-Sample Test	
D = max |F1 - F2|	0.1389
Asymptotic Pr > D	0.9931
Exact Pr >= D	.
D+ = max (F1 - F2)	0.0667
Asymptotic Pr > D+	0.9192
Exact Pr >= D+	.
D- = max (F2 - F1)	0.1389
Asymptotic Pr > D-	0.6939
Exact Pr >= D-	.

TABLE 3.6 – Statistiques de Kolmogorov-Smirnov pour tests bilatéral et unilatéral - Distribution des notes en session 1 du M1 ESA selon la provenance des étudiants

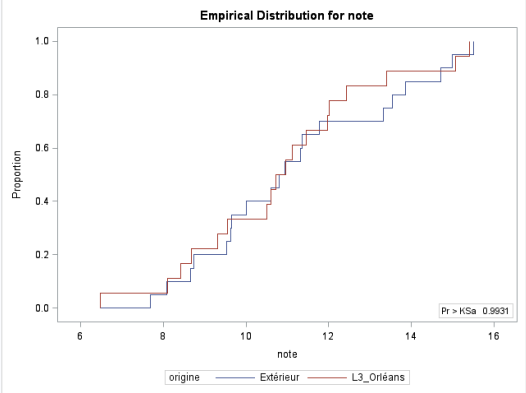


FIGURE 3.3 – Fonctions de répartition empiriques - Distribution des notes en session 1 du M1 ESA selon la provenance des étudiants.

Toujours en lien avec les statistiques de type *EDF*, la procédure fournit également le graphe des densités empiriques des deux échantillons, graphique reproduit dans la figure 3.3, qui permet éventuellement de repérer sur quels segments de notes les écarts observés sont les plus importants, par exemple ici entre 12 et 14/20. Attention cependant aux interprétations trop rapides. Ainsi, dans cet exemple, toutes les statistiques concluent à la non significativité des écarts. L'output continue par l'affichage d'informations relatives au test de Cramer-Von Mises, avec notamment les valeurs des statistiques CM et CM_a de la table 3.7 qui, sauf preuve du contraire, ne servent à rien.

La dernière statistique donnée par la procédure est celle de Kuiper, K et sa valeur asymptotique K_a donnée par l'équation 3.23. Comme on peut le constater dans la table 3.8, la conclusion de ce test est conforme à tout ce que nous avons obtenu jusqu'alors, à savoir le non rejet de l'hypothès nulle : les notes des deux catégories d'étudiants seraient tirées dans la même distribution, et aucun ne serait avantagé, ou désavantagé, en fonction de sa provenance.

Cramer-von Mises Statistics (Asymptotic)			
CM	0.000769	CMa	0.029240

TABLE 3.7 – Statistiques de Cramer-von Mises - D istribution des notes en session 1 du M1 ESA selon la provenance des étudiants

Kuiper Two-Sample Test (Asymptotic)					
K	0.205556	Ka	0.632687	Pr > Ka	0.9996

TABLE 3.8 – Statistiques de Kuiper - Distribution des notes en session 1 du M1 ESA selon la provenance des étudiants

3.3.5 Exemple 2 : le paradoxe de Simpson - analyse stratifiée et contrôle de la composition des échantillons

Le paradoxe de Simpson tient au fait que des écarts observés entre deux distributions peuvent s'inverser ou disparaître lorsque les échantillons sont combinés. La raison de la contradiction apparente est celle de l'absence de contrôle pour les effets d'une variable tierce sur celle qui est étudiée. Par exemple, dans le domaine médical, cela peut concerner une mesure de l'efficacité d'un traitement sur deux échantillons de personnes dont la composition par âge est différente. Dans une étude sur les salaires d'insertion de deux formations, cela peut être de conclure en faveur de l'une alors que les répartitions entre hommes et femmes sont différentes entre les filières et qu'existe une discrimination salariale en défaveur de celles-ci sur le marché du travail, etc...
Pour illustrer ce paradoxe, imaginons deux formateurs responsables chacun d'un groupe. Le premier, A est constitué de 15 femmes et 5 hommes, le second, B, de 5 femmes et 15 hommes. Faisons l'hypothèse que dans une évaluation commune, les femmes obtiennent toujours un score de 15, et les hommes un score de 10 quelqu'ait été leur formateur : la performance des deux formateurs est donc strictement identique. Pour autant, si on regarde le groupe A, la moyenne obtenue sera $(15 * 15 + 10 * 5)/20 = 13.75$, et pour le groupe "B" de $(15 * 5 + 10 * 15)/20 = 11.25$. Le formateur du groupe "A" serait au regard de la moyenne des élèves déclaré vainqueur. Un échantillon mélangé confirmerait ce résultat, ce formateur alignant 15 notes égales à 15 et 5 notes égales à 10 contre 5 notes de 15 et 15 notes égales à 10 pour le formateur "B" : ses scores occuperaient les rangs les plus élevés et un test de Wilcoxon ou un test de médiane rejetterait l'hypothèse d'égalité des distributions en sa faveur. Toute la question est d'apprécier l'impact de la composition selon le sexe des deux groupes sur le classement des formateurs. Notez qu'on n'est pas intéressé par la mesure de l'impact du sexe sur le score : il ne s'agit pas d'expliquer pourquoi les femmes ont 15 et les hommes ont 10, on veut simplement contrôler de ses effets pour répondre à la question qui nous intéresse [14]. C'est ce que va autoriser l'analyse stratifiée mise oeuvre par la commande `strata` de `npar1way`.

Afin d'illustrer la méthode, nous ne donnerons dans cet exemple que les résultats du test de Wilcoxon, sachant que tous les autres tests conduisent aux

14. En ce sens, la composition par sexe des groupes est ce qu'on appelle un facteur de confusion.

Wilcoxon Scores (Rank Sums) for Variable score Classified by Variable groupe					
groupe	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	20	510.0	410.0	32.025631	25.50
2	20	310.0	410.0	32.025631	15.50
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	510.0000
Normal Approximation	
Z	3.1069
One-Sided Pr > Z	0.0009
Two-Sided Pr > |Z|	0.0019
t Approximation	
One-Sided Pr > Z	0.0018
Two-Sided Pr > |Z|	0.0035
Z includes a continuity correction of 0.5.	

TABLE 3.9 – Exemple 2 - statistiques de Wilcoxon sans stratification

mêmes conclusions.

Soit dans un premier temps une comparaison des distributions sans correction. L'étape `data` suivante créée la table des données de l'exemple, l'appel de la proc `npar1way` qui suit réclamant la statistique de Wilcoxon usuelle :

```
data formateurs;
input score sexe $ groupe effectif;
cards;
15 f 1 15
10 h 1 5
15 f 2 5
10 h 2 15
;
run;
proc npar1way data=formateurs wilcoxon;
class groupe;
freq effectif;
var score;
run;
```

On obtient alors les résultats de la table 3.9. Comme attendu, la somme des rangs des scores issus du premier groupe est beaucoup plus élevée que celle afférent au second groupe et surtout nettement plus élevée que la somme espérée sous $H0$. Les probabilités critiques confirment la significativité de l'excès en autorisant sans ambiguïté le rejet de l'égalité des distributions en faveur d'une dominance de la répartition des scores du premier animateur.

Maintenant, sachant que la répartition entre hommes et femmes des deux groupes est dissemblable, nous pouvons essayer de tenir compte de cette différence en réalisant une étude stratifiée. La variable de stratification est celle dont les modalités vont définir, au sein de nos échantillons de base, des sous-groupes que l'on soupçonne hétérogènes. Dans le présent exemple, c'est évidemment la variable `sexe` avec ses deux modalités "h" et "f" qui va être retenue via une

Wilcoxon Scores by Strata for score						
Classified by groupe						
Stratum	N Obs	Class N Obs	Statistic	Expected	Std Dev	Mean
1	20	15	157.50	157.50	0.0	10.50
2	20	5	52.50	52.50	0.0	10.50

Stratified Wilcoxon (Van Elteren) Test for score							
Classified by groupe							
N Obs	N Strata	Statistic	Expected	Std Dev	Z	Pr < Z	Pr > |Z|
40	2	10.00	10.00	0.0	0.0000	0.5000	1.0000

TABLE 3.10 – Exemple 2 - statistiques de Wilcoxon avec stratification

commande `strata sexe`.
La démarche mise en oeuvre est la suivante : la statistique réclamée, ici Wilcoxon, sera calculée sur chacune des strates. Nous aurons donc une somme des rangs évaluée sur les scores des hommes, et une somme des rangs évaluée sur les scores des femmes [15]. La statistique stratifiée de Wilcoxon est simplement la moyenne pondérée par les effectifs des strates de ces diverses statistiques. Une *z-transform* est finalement opérée qui va permettre le calcul d'une probabilité critique asymptotique sur une distribution gaussienne standardisée. Ici, sur les deux strates hommes et femmes, la somme des rangs observée sera naturellement égale à la somme espérée sur *H0* puisque les scores au sein d'une strate sont les mêmes quel que soit le formateur considéré. En conséquence la statistique stratifiée sera aussi égale à sa valeur espérée sous la nulle et nous devrions obtenir une probabilité critique égale à l'unité, signalant qu'une fois pris en compte les effectifs différents d'hommes et de femmes au sein de nos deux groupes, il n'existe plus aucune différence significative entre les deux distributions de scores.

Pour vérifier cela, il suffit d'ajouter la commande `strata sexe` aux lignes de commandes précédentes, soit :

```
proc npar1way data=formateurs wilcoxon;
class groupe;
freq effectif;
var score;
strata sexe;
run;
```

Les résultats de la table 3.10 sont alors affichés. On obtient bien une *z-transform* exactement égale à zéro, évidemment non significative contrairement à la statistique non stratifiée de la table 3.9.

15. Une option `ranks=stratum` ou `ranks=overall` précise si les rangs en question doivent être calculés au sein d'une strate, par exemple, la somme des rangs des scores des hommes du premier groupe, ces rangs étant construits en considérant uniquement les scores mélangés des hommes, avec `ranks=stratum`, ou bien la somme des rangs des scores des hommes du premier groupe, ces rangs étant construits en considérant les scores mélangés des hommes et des femmes.

Chapitre 4

Tests sur plus de deux échantillons

On est maintenant en présence d'au moins trois échantillons constitués d'individus indépendants et la question posée est celle de la similarité des distributions dont ils sont issus. Le plus fréquemment, les échantillons correspondent aux catégories d'une variable qualitative, par exemple *hommes* versus *femmes*, ou encore aux découpages d'un échantillon par tranches d'âges, de revenus, de catégories socio-professionnelles, etc...

Dans le cadre des tests paramétriques cette problématique conduit généralement à la mise en oeuvre d'une analyse de la variance se terminant par le calcul d'un F de Fisher avec une hypothèse nulle qui est celle de l'égalité des espérances au sein de chaque sous-groupe avec des conditions de validité portant sur la normalité des observations et leur homoscédasticité.

Lorsque ces conditions ne sont pas remplies, une alternative possible est évidemment de faire appel à un test non paramétrique. Parmi ceux-ci, le plus utilisé est celui de Kruskal-Wallis qui généralise le test de Wilcoxon vu au chapitre précédent. Pour ce test, et ses variantes, l'hypothèse alternative est simplement la négation de l'hypothèse nulle. Nous présenterons également le test de Jonckheere-Terpstra, qui se distingue par la formulation de l'alternative : on va poser que sous H1, la médiane d'une variable croît (ou décroît) lorsqu'un certain ordre est retenu sur les échantillons. Un exemple d'application est le suivant : supposez que l'on fasse passer la même épreuve à des élèves de DEUG1, de DEUG2, de Licence, de Master 1 et de Master 2. L'alternative ne serait pas seulement une différence dans les notes moyennes des promotions d'étudiants, mais une augmentation de celles-ci lorsque l'on passe d'une catégorie à l'autre en respectant l'ordre indiqué [1]. L'intérêt du test de Jonckheere-Terspstra vient de ce que si H1 est vraie, sa puissance est supérieure à celle du test de Kruskal-Wallis qui retient une hypothèse alternative plus générale : il devrait rejeter plus souvent l'hypothèse nulle lorsqu'elle est fausse [2].

1. C'est du moins ce que l'on peut raisonnablement espérer.

2. C'est d'ailleurs ce que vous connaissez déjà avec le test paramétrique de Student : Soit H0 : $\theta = \theta_0$ et deux alternatives, la plus générale H1 : $\theta \neq \theta_0$, et la plus restrictive H1 : $\theta > \theta_0$. Sous cette dernière, le risque de première espèce est concentré sur la partie droite de la distribution, sous H1 : $\theta \neq \theta_0$, c'est seulement sa moitié. En conséquence, la valeur critique est moins élevée pour la dernière que pour la première et on est conduit à rejeter H0 plus souvent si H1 : $\theta > \theta_0$ est vraie.

4.1 Le test de Kruskal-Wallis et ses variantes

Ces tests vont réaliser une analyse de la variance sur des transformées (scores) des observations initiales. Si on est en présence de r catégories et si on note T_i le total des scores au sein de la $i^{\text{ème}}$ catégorie, $i = 1, \ldots, r$, la statistique de test est donnée par :

$$C = \left(\sum_{i=1}^{r} (T_i - E_0(T_i))^2 / n_i \right) / S^2 \tag{4.1}$$

où $T_i = \sum_{j=1}^{n} I_{(ij)} s(o_j)$, $s(o_j)$ est le score de la $j^{\text{ème}}$ observation dans l'échantillon mélangé [3] et $I_{(ij)}$ une indicatrice valant 1 si o_j est dans la catégorie i, 0 sinon. $E_0(T_i)$ est l'espérance du total des scores dans la catégorie i sous l'hypothèse nulle donnée par :

$$E_0(T_i) = \frac{n_i}{n} \sum_{j=1}^{n} s(o_j) \tag{4.2}$$

où n est la taille de l'échantillon mélangé, *i.e.*, $n = \sum_{i=1}^{r} n_i$.

Enfin, S^2 est la variance estimée de la série des scores :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (s(o_j) - \bar{s})^2, \text{ et } \bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(o_j) \tag{4.3}$$

On sait qu'asymptotiquement la statistique C est, sous la nulle, la réalisation d'un chi2 à $r - 1$ degrés de liberté.

4.1.1 Le test de Kruskal-Wallis

C'est le plus populaire. Il reprend les scores de Wilcoxon, i.e. $s(o_j) = R_j$ où R_j est le rang de la $j^{\text{ème}}$ observation dans l'échantillon mélangé. Il suffit d'ailleurs du mot clef `WILCOXON` pour qu'avec plus de deux sous-échantillons en jeu la procédure `NPAR1WAY`, calcule ce test de Kruskal-Wallis. On peut illustrer les calculs avec l'exemple suivant :

$$E_1 = \{1, 2, 3\}$$
$$E_2 = \{4, 5, 6\}$$
$$E_3 = \{7, 8, 9\}$$

3. En cas d'égalité des scores, la procédure classe les observations par ordre croissant, attribue un rang à chaque observation, calcule des scores sur la base de ces rangs puis attribue la moyenne de leurs scores aux observations concernées.

Il vient :

$$\sum_{j=1}^{9} s(o_j) = \sum_{j=1}^{9} o_j = 45$$

$$T_1 = 6, T_2 = 15, T_3 = 24$$

$$E_0(T1) = E_0(T2) = E_0(T3) = \frac{3}{9} \times 45 = 15$$

$$\bar{s} = 5, S^2 = \frac{16+9+4+1+0+1+4+9+16}{8} = 7.5, \text{et finalement}$$

$$C = \frac{\left((6-15)^2 + (24-15)^2\right)/3}{7.5} = \frac{54}{7.5} = 7.2$$

qui sous l'hypothèse de distributions *similaires* des trois échantillons est la réalisation (asymptotiquement) d'un $\chi^2(2)$ ce qui conduit à une probabilité critique de 2.73% et donc un rejet de la nulle au seuil de risque de 5%.

Les interprétations erronées des conclusions du test

La difficulté avec ce test vient de son interprétation : il faut savoir ce que signifie ici le terme "similaire". Souvent H0 est entendue comme signifiant que les échantillons ont même moyenne, ou encore les échantillons ont même médiane. Or si on considère l'équation (4.1) on remarque aisément que, pour des échantillons de même taille, C dépend des écarts entre les rangs moyens des différents échantillons.

Dans son Handbook of Biological Statistics, John H. McDonald [4] met en garde contre ces interprétations erronées du test de Kruskal-Wallis et illustre la difficulté au moyen des données reproduites dans la table 4.1 afférentes à trois échantillons de 18 observations chacun :

Les rangs moyens de ces trois groupes, égaux respectivement à 20.39, 27.50 et 34.61 sont suffisamment différents pour que la statistique de kruskal-Wallis associée (=7.36) ait une probabilité critique de 2.53% ce qui conduit, au seuil de 5%, au rejet de l'hypothèse nulle d'égalité de ces rangs moyens. Pour autant, on peut vérifier que les moyennes et les médianes des observations sont identiques dans les 3 groupes (respectivement 43.50 et 27.50).

4.1.2 Quelques variantes du test de Kruskal-Wallis

Identifier les scores aux rangs des observations comme le font les tests de Wilcoxon et de Kruskal-Wallis est un cas particulier de création de scores et on peut naturellement envisager d'autres solutions. En voici quelques unes parmi celles qu'offre la procédure NPAR1WAY.

Les scores médians

La règle de construction des scores est la suivante :

4. McDonald, J.H. 2014. Handbook of Biological Statistics, 3rd ed. Sparky House Publishing, Baltimore, Maryland. Le handbook est également disponible à l'adresse http ://biostathandbook.com

groupe1	groupe2	groupe3
1	10	19
2	11	20
3	12	21
4	13	22
5	14	23
6	15	24
7	16	25
8	17	26
9	18	27
46	37	28
47	58	65
48	59	66
49	60	67
50	61	68
51	62	69
52	63	70
53	64	71
342	193	72

TABLE 4.1 – Données d'exemple reprises à J.H. McDonald

$$s(R_j) = \begin{cases} 1 \text{si } R_j > (n+1)/2 \\ 0 \text{ sinon} \end{cases} \tag{4.4}$$

Dans l'aide de la procédure, l'utilisation de ces scores médians est préconisée pour des distributions symétriques et à queues épaisses. Si la question que l'on se pose est celle de l'égalité des médianes des différentes distributions, elle est préférable à la statistique de Kruskall-Wallis. Ainsi, sur le jeu de données précédent, on obtient une statistique de test nulle et évidemment une probabilité critique de 100%. Pour l'obtenir, il suffit de remplacer `wilcoxon` par le mot-clef `median` dans l'appel de la procédure.

Les scores de Van der Waerden

Ils sont donnés par

$$s(R_j) = \Phi^{-1}\left(R_j/(n+1)\right) \tag{4.5}$$

et leur emploi est conseillé lorsque les données sont des réalisations de gaussiennes. Vous savez cependant que dans ce cas un test paramétrique est généralement préférable. Pour le calcul d'une statistique fondée sur ces scores on doit faire apparaître les mot-clefs `normal` ou `VW`

Vous trouverez dans l'aide de NPAR1WAY plusieurs autres scores plus ou moins optimaux selon les configurations. En pratique, on déconseille à moins d'avoir de bonnes justifications, de choisir une statistique "exotique" : préférez Kruskal-Wallis et confirmez éventuellement vos conclusions avec les scores médians et les scores de Van der Waerden. Notez enfin qu'il est possible de

demander la calcul de la probabilité critique exacte via la commande `exact`, utilisée par prudence avec son option `/maxtime=` .

4.2 Test de Jonckheere-Terpstra pour alternative ordonnée

Ce test est utile lorsque, disposant de $r > 2$ échantillons $E_1, E_2, \ldots, E_r$ indépendants, l'hypothèse nulle et son alternative peuvent s'écrire :

H0 : $\theta_1 = \theta_2 = \ldots \theta_r$ versus
H1 : $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_r$ ou H1 : $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_r$ avec au moins une inégalité stricte (< ou >) dans ces dernières listes.

où θ_i est la médiane du $i^{\text{ème}}$ échantillon. Dans la mise en oeuvre du test, l'ordre de prise en compte des échantillons importe et l'utilisateur doit être en mesure de spécifier un ordre compatible avec H1. Dans l'exemple donné en introduction, il faudra que le premier échantillon de la liste soit constitué par les étudiants de DEUG1, le second par ceux de DEUG2,...le dernier par les inscrits en M2. Une instruction `ORDER` est disponible afin d'assurer cette cohérence

On considère les indices i, j tels que $i = 1, \ldots, r-1$ et $j = i+1, \ldots r$. Soit $J1_i$ le nombre d'observations des E_j supérieures aux observations de E_i et $J2_i$ le nombre d'observations égales dans E_i et les E_j. Soit enfin $J = \sum_i J1_i + \frac{1}{2} J2_i$. Sous l'hypothèse nulle, cette statistique J vérifie :

$$E_0(J) = \frac{n^2 - \sum_{i=1}^{r} n_i^2}{4} \text{ et,}$$
$$V_0(J) = A/72 + B(36n(n-1)(n-2) + C/(8n(n-1)), \text{ avec}$$

$$A = n(n-1)(2n+5) - \sum_i n_i(n_i-1)(2n_i+5) - \sum_j n_j(n_j-1)(2n_j+5)$$
$$B = \left(\sum_i n_i(n_i-1)(n_i-2)\right)\left(\sum_j n_j(n_j-1)(n_j-2)\right)$$
$$C = \left(\sum_i n_i(n_i-1)\right)\left(\sum_j n_j(n_j-1)\right)$$

SAS calcule une z-transform : $z = (J - E_0(J)) / \sqrt{V_0(J)}$ et affiche les probabilités critiques bilatérale et unilatérale de cette gaussienne, sachant que cette dernière est évaluée comme la probabilité qu'a une gaussienne centrée-réduite d'être

— supérieure à z si $z > 0$, avec alors H1 : $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_r$,
— inférieure à z si $z < 0$ avec alors H1 : $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_r$

et la présence d'au moins une inégalité stricte (< ou >) sous H1.

	réponses			
touchées	10	12	14	16
heurtées	12	18	20	22
fracassées	20	25	27	30

TABLE 4.2 – vitesse estimée en fonction du verbe employé dans la question

Un exemple d'application

Pour cet exemple, nous reprenons les données de la page Wikipédia consacrée au test de Jonckheere-Tepstra. Elle est relative à une expérimentation menée par Loftus et Palmer : on montre à 12 individus le film d'une collision entre deux voitures. Les personnes sont ensuite réparties en 3 groupes et on demande

— aux personnes du premier groupe, à quelle vitesse roulaient les voitures lorsqu'elles se sont touchées ?
— aux personnes du deuxième groupe, à quelle vitesse roulaient les voitures lorsqu'elles se sont heurtées ?
— aux personnes du troisième groupe, à quelle vitesse roulaient les voitures lorsqu'elles se sont fracassées ?

On soupçonne que le verbe utilisé dans la question va influencer l'estimation de la vitesse de sorte que plus ce verbe implique de violence et plus la vitesse estimée devrait être élevée. Dans ce cas, non seulement les médianes devraient être différentes, mais elles devraient augmenter uniformément lorsque l'on passe des réponses du groupe 1 à celles du groupe 2 puis à celles du groupe 3, *i.e.* sous H1 on a $\theta_1 < \theta_2 < \theta_3$.
Les réponses obtenues sont présentées dans la table 4.2

Le calcul de la statistique de Jonckheere-Terpstra s'effectue sous SAS avec la procédure `Freq` et l'option `jt` de la commande `table`. L'important est que les modalités créant les échantillons sur lesquels porte la relation d'ordre apparaissent en ligne et selon l'ordre désiré dans le tableau croisé. Dans notre exemple, nous pouvons ainsi soumettre les lignes suivantes :

```
data Ex_JT;
  input verbe $ vitesse;
  cards;
touché 10
touché 12
touché 14
touché 16
heurté 12
heurté 18
heurté 20
heurté 22
fracassé 20
fracassé 25
fracassé 27
fracassé 30
;
run;
```

```
proc freq data=Ex_JT order=data;
  tables verbe*vitesse / jt;
run;
```

Dans cet exemple, nous avons donc 3 classes : C1 associé à "touché", C2 à "heurté", C3 à "fracassé", l'ordre de ces classes étant défini par l'ordre supposé des médianes sous H1. Pour obtenir la statistique JT, on va comparer chaque classe avec celles qui lui sont "supérieures", ici, C1 avec C2 et C3, puis C2 avec C3. Pour chaque observation d'un échantillon donné, on compte le nombre d'observations qui lui sont supérieures dans les échantillons "supérieurs". Ainsi à la première observation de C1, 10, correspond un effectif de 8 (les 4 observations de C2 et les 4 observations de C3), avec la seconde, 12, nous trouvons 7 observations qui lui sont supérieures dans C2 et C3, et une observation de même montant. En menant le même raisonnement pour toutes les observations de C1 et C2, il vient :

$$\text{Pour C1 } : J1 = 8 + 7 + 7 + 7 = 29, J2 = 1$$
$$\text{Pour C2 } : J1 = 4 + 4 + 3 + 3 = 14, J2 = 1$$

et au final JT=29 + 14 + $\frac{1}{2}(1 + 1) = 44$. On ne reproduit pas les calculs plus fastidieux donnant l'espérance et la variance de cette statistique sous la nulle permettant le calcul de la z-transform. La procédure `freq` renvoie la table 4.3. La valeur positive de z associée à une faible probabilité critique permet de rejeter l'hypothèse d'une médiane commune et de conclure en faveur de l'hypothèse avancée, à savoir une augmentation de la vitesse estimée en fonction de la violence implicitement contenue dans le verbe employé.
Une remarque sur cet exemple : notez l'emploi de l'instruction `order=data` qui a imposé que les calculs se fassent selon l'ordre désiré sur les modalités de la variable "verbe", modalités mises en ligne dans le tableau croisé. Vous devez comprendre qu'en l'absence de cette instruction, SAS aurait utilisé l'ordre lexicographique et nous aurions alors eu en ligne les modalités successives suivantes "fracassé", "heurté" puis "touché" conduisant à l'hypothèse alternative H1 : $\theta_1 > \theta_2 > \theta_3$. Vous pouvez vérifier par vous-mêmes que la conclusion est inchangée, seul change le signe de la variable z qui devient négative [5].

Enfin, en cas de doute sur l'adéquation de la distribution asymptotique de la z-transform, il est possible de réclamer le calcul de la probabilité critique exacte afférente au test de Jonkheere-Tepstra. Il suffit pour cela de faire apparaître dans la `proc freq` la commande `exact jt / maxtime=..;`

5. Remarquez que si sur les données de cet exemple on demande le calcul de la statistique de Kruskal-Wallis via les commandes

```
proc npar1way data=EX_JT wilcoxon;
class verbe;
var vitesse;
run;
```

on obtient une statistique égale à 7.62 et une probabilité critique de 2.2 pour cent que l'on peut comparer à la bilatérale de Jonckheere-Terpstra, égale à 3.3 pour mille, signalant ainsi la plus grande puissance de JT relativement à KW.

Statistics for Table of verbe by vitesse

Jonckheere-Terpstra Test	
Statistic	44.0000
Z	2.9392
One-sided Pr > Z	0.0016
Two-sided Pr > |Z|	0.0033

Sample Size = 12

TABLE 4.3 – Exemple de calcul de la statistique de Jonckheere-Tepstra

Chapitre 5

Tests d'indépendance ou d'homogénéité

5.1 Le χ^2 de Pearson

Il s'agit simplement d'adapter les énoncés et remarques faits dans la présentation de cette statistique lors de son application au test de validation d'une distribution théorique sur un seul échantillon.

5.1.1 Test d'homogénéité d'échantillons

Dans le cas d'un test d'homogénéité, les observations des échantillons sont réparties en k classes. Celles-ci peuvent être définies soit par une valeur discrète, ordinale ou nominale, soit par regroupement des observations d'une variable continue. On suppose que ces regroupements sont complets : toute observation appartient à une classe et une seule. Avec deux échantillons on observera ainsi un tableau croisé semblable à celui représenté dans la table 5.1, ou $n_{i,j}$ est l'effectif des observations de l'échantillon i ($i = X, Y$) dans la classe j, et n_{c_i} celui de la classe i. Soit $p_{x,i}$ et $p_{y,i}$ les probabilités respectives qu'une réalisation de chacune des deux aléatoires appartienne à la classe i. L'hypothèse nulle est celle d'absence de différence entre les échantillons quant à la répartition des individus dans les classes. C'est donc un test d'homogénéité des échantillons au regard de la variable définissant ces classes. Si les deux échantillons sont constitués de réalisations indépendantes tirées dans la même loi alors la probabilité d'appartenance à une classe quelconque ne varie pas selon l'échantillon et l'hypothèse nulle se formule alors comme $H0 : p_{x,i} = p_{y,i}, i = 1, 2, \ldots, k$. Il s'agit naturellement d'une hypothèse jointe, l'alternative étant que l'une des classes n'ait pas la même probabilité de survenue dans X et Y. Sous cette hypothèse nulle, il n'y a pas de raison de distinguer les échantillons, et le meilleur estimateur de la probabilité d'appartenance à une classe i est évalué dans l'échantillon joint selon : Pr[appartenance à la classe i] $= \frac{n_{c_i}}{n}$. En conséquence, les effectifs espérés sous la nulle dans la classe i pour les observations en provenance de X et Y sont

Échantillon	classes 1	classe 2	...	classe k	Total
X	$n_{x,1}$	$n_{x,2}$	...	$n_{x,k}$	n_x
Y	$n_{y,1}$	$n_{y,2}$	...	$n_{y,k}$	n_y
Total	n_{c_1}	n_{c_2}	...	n_{c_k}	n

TABLE 5.1 – exemple de tableau croisé avec 2 échantillons

respectivement égaux à $e_{x,i} = \frac{n_{c_i}}{n} \times n_x$ et $e_{y,i} = \frac{n_{c_i}}{n} \times n_y$ [1].

La statistique de test est construite sur la totalité de ces écarts entre effectifs observés et espérés selon :

$$Q_P = \sum_{i=1}^{k} \frac{(n_{x,i} - e_{x,i})^2}{e_{x,i}} + \frac{(n_{y,i} - e_{y,i})^2}{e_{y,i}} \tag{5.1}$$

Plus les écarts sont importants et plus la statistique sera élevée, augmentant ainsi la probabilité d'un rejet de l'hypothèse nulle. On sait que cette statistique va être distribuée asymptotiquement selon un chi-deux, mais pour mettre en application ce résultat, il faut évidemment identifier son nombre de degrés de liberté, c'est à dire le nombre de $\chi^2(1)$ indépendants présents dans les $2 \times k$ utilisés dans la construction de Q_P. Dans ce type d'exercice, les effectifs des échantillons X et Y sont souvent des constantes déterminées avant l'analyse statistique proprement dite. En d'autres termes, les quantités n_x et n_y sont connues et fixes. En conséquence si on considère une ligne de la table relative à l'un ou l'autre des échantillons, seuls les effectifs de $k-1$ classes peuvent être indépendants : soit $S_{x,k-1}$ la somme des effectifs de $k-1$ classes prises au hasard pour l'échantillon X, alors nécessairement l'effectif de la dernière est $n_x - S_{x,k-1}$. A ce stade, il semble donc que le nombre de degrés de liberté des lignes relatives à X d'une part et Y d'autre part soit $k-1$ et non pas k. Maintenant, on sait que sous $H0$, la distinction entre X et Y ne joue pas. Cela signifie que, sous $H0$ la seule information qui compte est la répartition des n individus, quelle que soit leur provenance entre les k classes : si on mélange les échantillons, on obtient alors la ligne de marge du tableau 5.1, et cette ligne est donc, sous $H0$ une donnée connue et fixée pour les découpages en échantillons de l'ensemble de la population. En d'autres termes, non seulement la colonne de marge mais aussi la ligne de marge de la table sont des données qui s'imposent aux effectifs des cases intérieures. Dans ces conditions, si on connaît $k-1$ effectifs de classe pour X alors on peut en déduire les $k-1$ effectifs de classe de l'échantillon Y. Par exemple, si $n_{x,1}$ est connu alors obligatoirement $n_{y,1} = n_{c_1} - n_{x,1}$. Au final, seuls $k-1$ effectifs des classes réparties entre les deux échantillons peuvent être indépendants et déterminent tous les autres lorsque les marges du tableau sont fixes : en conséquence, le χ^2 précédent est à $k-1$ degrés de liberté [2]. L'appel de ce test d'homogénéité s'effectue naturellement toujours au sein de la commande `TABLE` au moyen de son option `chisq`.

1. On peut évidemment intervertir lignes et colonnes, le raisonnement est parfaitement symétrique puisque $e_{x,i} = \left(\frac{n_{c_i}}{n}\right) \times n_x = \left(\frac{n_x}{n}\right) \times n_{c_i}$.

2. En anticipant un peu, on peut retenir un résultat général, à savoir que dans un tableau $r \times c$ de r lignes et c colonnes à marges fixes, le nombre de degrés de liberté du χ^2 de Pearson est égal à $(r-1)(c-1)$. Ici $r = 2, c = k$ et on retrouve bien avec cette règle le $k-1$ du texte.

opinion	Femmes	Hommes	Total
Pas du tout d'accord	2	2	4
Plutôt pas d'accord	7	6	13
Plutôt d'accord	7	14	21
Tout à fait d'accord	0	4	4
Ensemble	16	26	42

TABLE 5.2 – Opinion d'étudiants sur l'affirmation "L'argent fait le bonheur !"

5.1.2 Test d'indépendance de variables

Pour un test d'indépendance entre deux variables, souvent nominales ou ordinales, une démarche identique à la précédente sera menée. Supposons ainsi que le tableau dénombre les réponses données à deux questions dont les réponses sont évaluées sur une échelle de Likert, par exemple "Très satisfait", "Satisfait", "Insatisfait", "Très insatisfait". L'hypothèse nulle est l'indépendance des réponses aux deux questions, l'alternative étant que la réponse faite par une personne à l'une des questions contient une information sur la réponse qu'elle donnera à l'autre question. En cas d'indépendance, la probabilité jointe est égale au produit des probabilités marginales, et donc la probabilité de prendre le choix i pour une des questions et le choix j pour l'autre est $p_{ij} = p_i p_j$. Cela signifie que, sous $H0$, l'effectif espéré au croisement des choix i et j est $e_{ij} = n \times p_{ij} = n \times p_i \times p_j = n \times \frac{n_{i.}}{n} \times \frac{n_{.j}}{n} = \frac{n_{i.} n_{.j}}{n}$, ce qui est bien l'équivalent de la définition des effectifs espérés donnée dans le test d'homogénéité des échantillons. Enfin, si on a les mêmes dimensions que précédemment [3], soit ici, en colonne une question à k modalités, et une question à 2 modalités en ligne, on aura également à la sortie la réalisation d'une khi-deux à $(k-1)$ degrés de liberté.

Dans les deux cas, les probabilités critiques asymptotiques sont donc calculées à partir d'une distribution $\chi^2(k-1)$. Il faut toutefois que des conditions de validité des résultats asymptotiques soient vérifiées, et il est conseillé en pratique d'avoir un effectif total, n, d'au moins 30 observations et une table telle qu'aucun effectif espéré ne soit inférieur à 5. Dans le cas contraire, il faut ou bien regrouper des classes, ou bien si possible augmenter la taille de ou des échantillons de travail. Il est également possible, via la commande `exact`, d'obtenir leurs probabilités critiques exactes, ce qui est naturellement utile lorsque n est faible, mais aussi lorsqu'il y a beaucoup d'observations égales entre elles ou bien encore lorsque les effectifs de certaines cases du tableau croisé sont faibles.

5.1.3 Exemple : L'argent fait le bonheur !

Nous avons demandé à 42 étudiants, 16 filles et 26 garçons s'ils étaient tout à fait d'accord, plutôt d'accord, plutôt pas d'accord ou pas du tout d'accord avec cette proposition. Leurs réponses sont présentées dans la table 5.2.

Il s'agit de savoir si le genre influence l'avis que l'on donne quant à cette affirmation. L'hypothèse nulle est que l'opinion est indépendante du sexe de la personne interrogée. La première étape consiste donc à construire les effectifs

3. i.e. un tableau $2 \times k$ correspondant à la répartition en k classes des observations de deux échantillons.

sous $H0$. Considérons par exemple l'opinion "Plutôt d'accord" qui est choisie par 21 étudiants sur 42. Sans distinction de sexe, la probabilité de cette opinion est donc de 0.5. Avec 16 filles et 26 garçons, sous l'hypothèse nulle on espère donc que 8 filles et 13 garçons devraient sélectionner cette réponse. On peut répéter ce raisonnement pour chaque avis possible. Plusieurs fréquences espérées étant inférieures à 5, nous ne sommes pas dans les conditions qui permettraient de se référer à la distribution asymptotique du khi-deux de Pearson. Deux solutions sont possibles : regrouper des catégories ou demander le calcul des probabilités critiques exactes.

1. Le programme ci-après réclame ces probabilités exactes [4].

```
proc format;
  value fopinion
    4 = "Tout à fait d'accord"
    3 = "Plutôt d'accord"
    2 = "Plutôt pas d'accord"
    1 = "Pas du tout d'accord"
;
  value fsexe
    0 = "Femmes"
    1 = "Hommes"
;
  value favis
    1 ="Oui"
    0 ="Non"
;
run;
data lovemoney;
  input sexe opinion effectifs;
  cards;
0 4 0
0 3 7
0 2 7
0 1 2
1 4 4
1 3 14
1 2 6
1 1 2
;
run;
proc freq data=lovemoney order=internal;
  table sexe*opinion / chisq expected;
  weight effectifs;
  format opinion fopinion.;
  format sexe fsexe.;
  exact chisq / maxtime=60;
run;
```

4. Pour information, dans la commande `tables` nous faisons apparaître l'option `Expected` qui demande l'affichage dans le tableau croisé des fréquences espérées sous la nulle.

Frequency Expected Percent Row Pct Col Pct	Table of sexe by opinion				
		opinion			
	sexe	Pas du tout d'accord	Plutôt pas d'accord	Plutôt d'accord	Tout à fait d'accord	Total
	Femmes	2 1.5238 4.76 12.50 50.00	7 4.9524 16.67 43.75 53.85	7 8 16.67 43.75 33.33	0 1.5238 0.00 0.00 0.00	16 38.10
	Hommes	2 2.4762 4.76 7.69 50.00	6 8.0476 14.29 23.08 46.15	14 13 33.33 53.85 66.67	4 2.4762 9.52 15.38 100.00	26 61.90
	Total	4 9.52	13 30.95	21 50.00	4 9.52	42 100.00

TABLE 5.3 – Statistiques du croisement des opinions avec le sexe

Statistic	DF	Value	Prob
Chi-Square	3	4.2714	0.2336
Likelihood Ratio Chi-Square	3	5.5968	0.1330
Mantel-Haenszel Chi-Square	1	3.2433	0.0717
Phi Coefficient		0.3189	
Contingency Coefficient		0.3038	
Cramer's V		0.3189	
WARNING: 63% of the cells have expected counts less than 5. (Asymptotic) Chi-Square may not be a valid test.			

Pearson Chi-Square Test	
Chi-Square	4.2714
DF	3
Asymptotic Pr > ChiSq	0.2336
Exact Pr >= ChiSq	0.2268

TABLE 5.4 – Statistiques de Pearson et probabilités critiques

On récupère alors les résultats de la table 5.3 détaillant les statistiques issues du croisement des réponses des Femmes et des Hommes aux quatre modalités de l'échelle de Likert retenue et ceux de la table 5.4 affichant notamment la valeur du χ^2 de Pearson et ses probabilités critiques asymptotiques et exactes. Avec 3 degrés de liberté et une valeur de 4.2714, nous ne pouvons pas rejeter l'hypothèse nulle d'indépendance entre les opinions et le sexe des répondants. Notez dans la sortie le message signalant que l'on est en-dessous du seuil de 5 individus minimum espérés dans certaines cases.

2. Regroupement des modalités.
   On ne va plus retenir que deux modalités pour les avis : soit favorable (regroupement de "Tout à fait d'accord" et "Plutôt d'accord"), soit défavorable (regroupement de "Pas du tout d'accord" et "Plutôt pas d'accord") à l'affirmation. Le programme suivant recombine les observations et teste l'indépendance dans la table 2×2 résultante.

```
data moneylove;
  input sexe avis effectifs;
  cards;
0 1 7
0 0 9
```

Frequency Expected Percent Row Pct Col Pct	Table of sexe by avis		
	avis		
sexe	Non	Oui	Total
Femmes	9 6.4762 21.43 56.25 52.94	7 9.5238 16.67 43.75 28.00	16 38.10
Hommes	8 10.524 19.05 30.77 47.06	18 15.476 42.86 69.23 72.00	26 61.90
Total	17 40.48	25 59.52	42 100.00

TABLE 5.5 – "L'argent fait le bonheur" - statistiques du croisement des opinions avec le sexe après regroupement des modalités initiales

Statistic	DF	Value	Prob
Chi-Square	1	2.6692	0.1023
Likelihood Ratio Chi-Square	1	2.6646	0.1026
Continuity Adj. Chi-Square	1	1.7163	0.1902
Mantel-Haenszel Chi-Square	1	2.6056	0.1065
Phi Coefficient		0.2521	
Contingency Coefficient		0.2444	
Cramer's V		0.2521	

Pearson Chi-Square Test	
Chi-Square	2.6692
DF	1
Asymptotic Pr > ChiSq	0.1023
Exact Pr >= ChiSq	0.1205

TABLE 5.6 – Statistiques de Pearson et probabilités critiques après regroupement

```
1 1 18
1 0 8
;
run;
proc freq data=moneylove order=internal;
  table sexe*avis / chisq;
  weight effectifs;
  format avis favis.;
  format sexe fsexe.;
  exact chisq / maxtime=60;
run;
```

La procédure affiche alors la table 5.5, où apparaissent encore les effectifs espérés sous $H0$. On respecte maintenant les conditions préconisées pour utiliser la distribution asymptotique du khi-deux de Pearson [5]. Les résultats des tests sont présentés dans la table 5.6.

Aux seuils de risque de 5% ou 10%, nous ne pouvons toujours pas

5. Pour mémoire, un effectif total d'au moins 30 individus et un minimum de 5 dans chaque case du tableau dans une table 2×2

rejeter l'hypothèse d'indépendance des avis et du sexe des enquêtés que ce soit avec la probabilité critique asymptotique ou la probabilité critique exacte.

5.2 Le cas des tableaux $R \times C$

Tout ce qui vient d'être présenté se généralise sans difficulté à des tableaux de dimensions quelconques $R \times C$ avec donc $R \geq 2$ et $C \geq 2$. Par exemple, la statistique de Pearson devient :

$$Q_P = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(o_{r,c} - e_{r,c})^2}{e_{r,c}} \tag{5.2}$$

ou $o_{r,c}$ et $e_{r,c}$ sont respectivement les effectifs observés et espérés sous $H0$ au croisement de la ligne r et de la colonne c. En l'absence d'association entre les lignes et les colonnes, cette statistique est asymptotiquement la réalisation d'un $\chi^2((R-1)(C-1))$.

5.3 Le test de Fisher exact

Créé par Fisher pour des tables 2×2, ce test s'applique en théorie lorsque les marges horizontale et verticale de la table sont fixes. Un avantage par rapport au khi-deux de Pearson dans cette configuration est que l'on peut calculer la probabilité critique exacte d'une statistique. N'ayant plus besoin de se référer à une distribution asymptotique, ce test est notamment employé lorsque les conditions préconisées pour atteindre cette distribution ne sont pas vérifiées, par exemple, lorsqu'on a moins de 5 individus espérés dans une ou plusieurs cases de la table. Le test a ensuite été étendu aux tables *RxC* quelconques par Freeman et Halton [6].

5.3.1 Présentation du test pour les tables 2x2

On est donc en présence d'un tableau semblable à celui ci :

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

où les effectifs $n_{1.}, n_{2.}, n_{.1}, n_{.2}$, et donc n sont des constantes, les seules aléatoires étant les fréquences $n_{ij}, i, j = 1, 2$. Dans ces conditions, il n'y a qu'une seule réalisation de libre : dès lors que l'une est connue, les trois autres sont entièrement déterminées. Ainsi, si on se donne n_{11}, alors $n_{12} = n_{1.} - n_{11}$, $n_{21} = n_{.1} - n_{11}$, $n_{22} = n_{2.} - n_{21}$. Formellement, il s'agit d'un tirage simultané, *i.e.* sans remise, de $n_{1.}$ individus parmi n et on s'intéresse à la variable aléatoire X dont la réalisation est le nombre n_{11} d'apparitions d'éléments ayant le caractère étudié sachant que

6. Freeman, G. H., and Halton, J. H. (1951), Note on an Exact Treatment of Contingency, Goodness of Fit, and Other Problems of Significance, *Biometrika* 38, 141-149.

l'effectif de ces éléments dans la population est $n_{.1}$. Cette variable X possède une distribution hypergéométrique de paramètres $(n, n_{1.}, n_{.1})$, et on a :

$$Pr(x = n_{11}) = \frac{\binom{n_{1.}}{n_{11}}\binom{n - n_{1.}}{n_{.1} - n_{11}}}{\binom{n}{n_{.1}}} = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!} \tag{5.3}$$

On peut donc évaluer la probabilité de la table observée sur les données de travail. La probabilité critique exacte, quant à elle, est la probabilité d'observer cette table ou toute table correspondant à une répartition plus extrême des individus. Pour la calculer, il suffit donc de construire les tables plus extrêmes, de calculer leurs probabilités et d'ajouter celles-ci à la probabilité de la table observée.
Par ailleurs, en gérant la construction des tables plus extrêmes, il est possible de calculer une probabilité critique unilatérale ou bilatérale. Par exemple, pour effectuer un test unilatéral à gauche, on va considérer comme étant plus extrêmes les tables qui ont un effectif n_{11} inférieur ou égal à celui observé. Pour un test unilatéral à droite, il suffira de considérer les tables pour lesquelles l'effectif n_{11} est supérieur ou égal à celui observé dans le tableau initial.

Pour illustrer la démarche, on repart du tableau 2×2 observé à l'exemple précédent qui répartissait en deux modalités ("Non" et "Oui") les opinions de deux échantillons de filles et de garçons quant à l'affirmation "l'argent fait le bonheur". Nous reprenons cette table observée en tête de la table 5.7, puis, dans les lignes suivantes, toutes les tables plus extrêmes dans l'optique d'un test unilatéral à droite. Les marges de ces tables coloriées en jaune sont par construction fixes. L'hypothèse nulle est que la table observée ne permet pas de rejeter l'indépendance des opinions au regard du sexe, l'hypothèse alternative étant que la propension des filles à répondre "non" est plus élevée qu'elle ne le serait sous cette hypothèse d'indépendance de l'opinion et du genre. La seconde colonne donne la probabilité de chacune de ces tables, probabilité évaluée au moyen de (5.3.1) . Ainsi, pour ce test unilatéral à droite on obtient :

$$\begin{aligned} \text{Probabilité critique exacte} &= 0.0702 + 0.0207 + 0.0039 + 0.0005 \\ &= 0.0953 \end{aligned} \tag{5.4}$$

résultat qui ne permet pas de rejeter la nulle au seuil de 5%.

Pour trouver la probabilité critique exacte du test bilatéral, $H0$: "les opinions sont indépendantes du sexe", versus $H1$: "les opinions des filles et des garçons sont différentes", il suffit d'ajouter à la probabilité précédente celles de tous les tableaux 2×2 ayant les mêmes marges fixes mais construits dans la direction opposée à celle prise pour le test unilatéral précédent, et dont la probabilité de survenue est inférieure ou égale à celle du tableau observé, *i.e.* qui correspondent à des répartitions au moins aussi extrêmes que celle observée sur les échantillons de travail. Les éléments de calculs sont donnés dans la table 5.8.

Seules les quatres dernières tables ont une probabilité inférieure à 0.0702.

table	$Pr[table]$
9 7 16 8 18 26 17 25 42	0.0702
10 6 16 7 19 26 17 25 42	0.0207
11 5 16 6 20 26 17 25 42	0.0039
12 4 16 5 21 26 17 25 42	0.0005
13 3 16 4 22 26 17 25 42	<0.0001
14 2 16 3 23 26 17 25 42	<0.0001
15 1 16 2 24 26 17 25 42	<0.0001
16 0 16 1 25 26 17 25 42	<0.0001

TABLE 5.7 – Test de Fisher exact - Exemple de calcul d'une probabilité critique unilatérale à droite.

table	$Pr[table]$
8 8 16 9 17 26 17 25 42	0.1579
7 9 16 10 16 26 17 25 42	0.2386
6 10 16 11 15 26 17 25 42	0.2430
5 11 16 12 14 26 17 25 42	0.1657
4 12 16 13 13 26 17 25 42	0.0743
3 13 16 14 12 26 17 25 42	0.0212
2 14 16 15 11 26 17 25 42	0.0036
1 15 16 16 10 26 17 25 42	0.0003
0 16 16 17 9 26 17 25 42	0.0001

TABLE 5.8 – Test de Fisher exact - Exemple de calcul d'une probabilité critique bilatérale.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	9
Left-sided Pr <= F	0.9749
Right-sided Pr >= F	0.0953
Table Probability (P)	0.0702
Two-sided Pr <= P	0.1205

TABLE 5.9 – Résultats du test de Fisher exact sur les données de l'exemple "l'argent fait le bonheur"

En conséquence,

$$\begin{aligned} \text{Probabilité critique exacte pour test bilatéral} &= 0.0953 + 0.0212 + 0.0036 \\ &\quad + 0.0003 + 0.0001 \\ &= 0.1205 \end{aligned} \tag{5.5}$$

On peut réaliser ce test de Fisher exact dans la procédure `freq` via le commande `exact fisher`. L'exemple précédent peut donc être reproduit avec les commandes suivantes :

```
data moneylove;
input sexe avis effectifs;
cards;
0 0 9
0 1 7
1 0 8
1 1 18
;
run;
ods html close;
ods html;
proc freq data=moneylove order=internal;
table sexe*avis / chisq expected;
weight effectifs;
format avis favis.;
format sexe fsexe.;
exact fisher;
run;
```

Les résultats sont donnés dans la table 5.9. et on y retrouve bien les probabilités critiques exactes afférentes aux versions unilatérale et bilatérale calculées *à la main*. Pour terminer, on peut encore noter que l'on conclut toujours, au seuil de 5%, au non rejet de l'hypothèse d'indépendance entre le sexe et les opinions.

5.3.2 Extension aux tables *RxC* quelconques

La loi hypergéométrique permet de calculer la probabilité de survenue d'une table à marges fixes de dimension quelconque. Il est donc possible de calculer la probabilité p_o de la table observée et la somme des probabilités de

Frequency Percent Row Pct Col Pct	Table of genre by vin			
	vin			
genre	blanc	rose	rouge	Total
femme	5 15.15 31.25 62.50	5 15.15 31.25 55.56	6 18.18 37.50 37.50	16 48.48
homme	3 9.09 17.65 37.50	4 12.12 23.53 44.44	10 30.30 58.82 62.50	17 51.52
Total	8 24.24	9 27.27	16 48.48	33 100.00

Fisher's Exact Test	
Table Probability (P)	0.0484
Pr <= P	0.4905

TABLE 5.10 – Résultats du test de Fisher exact sur les données de l'exemple "genre et préférence de vin"

toutes les tables plus extrêmes que celle-ci, tables que l'on identifie par leur probabilité de survenue inférieure ou égale à p_o. Comme précédemment cette somme constitue la probabilité critique du test bilatéral de Fisher exact. Une valeur inférieure au seuil de risque α mène à la conclusion d'une association entre les lignes et les colonnes de la table. Notez encore que pour les tables de dimensions supérieures à 2x2, seul un test bilatéral est réalisable.
Pour illustration, supposez que l'on ait demandé à 17 hommes et 16 femmes de donner leur préférence en ce qui concerne le vin : rouge versus rosé versus blanc. Le programme suivant vous donne les données et les commandes utilisées, la table 5.10 présente les résultats obtenus. On ne pourrait pas, avec cet échantillon qui ne se prétend pas représentatif, l'hypothèse d'indépendance des préférences et du genre. Notez aussi que la faiblesse des effectifs dans certaines cases du tableau fait douter de l'adéquation du test de Chi-2 standard pour traiter cet exemple.

```
data prefer ;
input genre vin effectif ;
cards ;
homme rouge 10
homme rose 4
homme blanc 3
femmes rouges 6
femmes rose 5
femme blanc 5
;
proc freq data=prefer ;
tables genre*vin ;
weight effectif ;
exact fisher ;
run ;
```

catégorie sociale	Hommes (0)	Femmes (1)
Ouvriers	n_{H1}	n_{F1}
Cadres et Professions intellectuelles	n_{H2}	n_{F1}
Professions intermédiaires	n_{H3}	n_{F3}
Employés	n_{H4}	n_{F4}

TABLE 5.11 – Exemple de configuration pour laquelle le test de Cochran-Armitage est adapté

5.4 Hypothèse alternative avec tendance : le test de Cochran-Armitage

Le test de Cochran-Armitage joue, relativement au test de Chi-2 de Pearson, le même rôle que le test de Jonckheere-Tepstra vis-à-vis du test de Kruskal-Wallis. Il s'agit de tester la dépendance entre une variable prenant deux valeurs et une variable à plusieurs modalités. Dans cette présentation on va supposer que cette dernière variable est mise en ligne : on est en présence d'une table $R \times 2$ contenant des effectifs et on va supposer que, sous l'hypothèse alternative, les proportions mesurées sur les lignes varient linéairement en fonction des R modalités lorsque celles-ci respectent un ordre que l'on se donne a-priori.
Par exemple, supposez que l'on dispose d'un échantillon d'hommes et de femmes découpé en fonction de la catégorie sociale selon l'ordre présenté dans la table 5.11.

Ce test peut être utilisé avec le jeu d'hypothèses suivant :

$H0$:Pr[être une femme | cat=ouvriers]=
Pr[être une femme | cat=cadres] =
Pr[être une femme | cat=professions intermédiaires]
Pr[être une femme | cat=employés], versus
$H1$:Pr[être une femme | cat=ouvriers]<
Pr[être une femme | cat=cadres] <
Pr[être une femme | cat=professions intermédiaires]<
Pr[être une femme | cat=employés].

Naturellement, sous H1 il est possible d'inverser le sens des inégalités et de passer de "<" à ">". Le test bilatéral correspondrait au cas où on ne connaît pas

a priori les sens des inégalités. Le H1 dans ce cas a pour expression :

$H1$:Pr[être une femme | cat=ouvriers]<
Pr[être une femme | cat=cadres] <
Pr[être une femme | cat=professions intermédiaires]<
Pr[être une femme | cat=employés].

ou

$H1$:Pr[être une femme | cat=ouvriers]>
Pr[être une femme | cat=cadres] >
Pr[être une femme | cat=professions intermédiaires] >
Pr[être une femme | cat=employés].

Dans cet exemple, on peut ainsi poser que sous l'alternative, la proportion de femmes augmente au fur et à mesure que l'on descend dans les lignes du tableau 5.11 [7] ce qui est évidemment plus restrictif que l'hypothèse alternative que retiendrait un test de Pearson. Comme pour le test de Jonckheere-Tespstra, si la présence d'une tendance dans l'évolution des proportions est vraie, alors la capacité de Cochran-Armitage à rejeter la nulle devrait être supérieure à celle du test de Pearson qui utilise comme alternative simplement la négation de H0. Plus précisément, les effectifs de la table $R \times 2$ [8], sont donc des réalisations de binomiales et, sous H1 on postule l'existence d'une dépendance linéaire de la probabilité d'une des réponses possibles, ici hommes ou femmes, en fonction des modalités ordonnées d'une autre variable, ici la catégorie sociale. Une fois ces modalités converties en score, H1 correspond alors à l'équation du modèle de probabilité linéaire suivante :

$$Pr[\text{réponse=1}] = \beta_0 + \beta_1 s_i \tag{5.6}$$

l'hypothèse nulle se traduisant par H0 : $\beta_1 = 0$.

Lorsque la variable de catégorisation est nominale, les scores associés à ses modalités sont simplement les numéros de ligne. Dans notre exemple, le score de la modalité "ouvriers" serait 1, celui de la modalité "cadres" serait 2, etc... Pour des variables numériques, la construction des scores est gérée par l'option `Scores=` de la commande `Tables`. Avec une variable de catégorisation numérique plusieurs possibilités sont offertes via l'option `score=`. Avec `score=table`, qui est la valeur de cette option par défaut, s_i est égal à la valeur de la i^{ème} modalité. Avec `score=rank`, s_i est égal à la moyenne des rangs qu'occupent les observations de la i^{ème} modalité, sachant que celles de la première occupent les rangs allant de 1 à $n_{1.}$, celle de la deuxième de $n_{1.} + n_{2.}$, etc... [9]. Deux autres possibilités existent : `scores=ridit` et `scores=moridit` qui réduisent les scores

7. Pour information, l'INSEE estime qu'en 2012 la proportion des femmes est de 19.6% chez les ouvriers, de 40.2% chez les cadres et Professions intellectuelles, de 51.2% dans les Professions intermédiaires et de 76.6% chez les Employés.

8. Naturellement le test peut être réalisé sur une table à 2 lignes et C colonnes, et la tendance porte alors sur la déformation des probabilités d'une ligne lorsqu'on se déplace d'une colonne à une autre. SAS va automatiquement adapter les calculs en fonction des dimensions de la table étudiée.

9. Par exemple, avec $n_{1.} = 3, n_{2.} = 2, n_{3.} = 3$, les scores renvoyés par `scores=rank` seront respectivement égaux à 2, 4.5 et 7.

des rangs précédents en les divisant respectivement par n ou $n + 1$.

Naturellement l'existence d'une tendance est associée à un ordre particulier des modalités de la variable de catégorisation. Pour définir l'ordre de prise en compte de ses modalités, on dispose comme pour le test de Jonckheere et Tepstra déjà vu, de l'option `order=` dans l'appel de `proc freq`. La statistique est donnée par [10]

$$T = \sum_{i=1}^{R} n_{i1}(s_i - \bar{s}) / \sqrt{p_{.1}(1 - p_{.1})s^2} \tag{5.7}$$

avec la proportion $p_{.1} = n_{.1}/n$ et

$$s^2 = \sum_{i=1}^{R} n_{i.}(s_i - \bar{s})^2 \tag{5.8}$$

Elle est réclamée par l'option `trend` dans la commande `Tables`. La procédure affiche alors la z-transform associée, la probabilité critique asymptotique unilatérale étant calculée comme $Pr[z > T]$ si $T > 0$ ou $Pr[z < T]$ si $T \leq 0$, et la bilatérale par $Pr[z > |T|]$.

Retenez encore qu'en cas de doute sur l'adéquation de la distribution asymptotique, le calcul des probabilités critiques exactes est également possible via la commande `exact trend / maxtime=`, et enfin qu'en cas de temps de calcul trop long, on peut recourir à une estimation de type bootstrap via l'option `mc` de cette même commande `exact`.

Exemple d'application

Pour cet exemple nous allons reprendre les données utilisées par Cochran. Des malades ont reçu un traitement soit à dose élevée, soit à dose faible. L'évolution de leur état est classée en 5 catégories : détérioration, stabilité, légère amélioration, amélioration modérée et amélioration marquée. Il s'agit d'étudier l'impact du dosage sur cette évolution. Les lignes ci-après créent la base.

```
data cochran;
input evolution :$20. dosage $ effectif;
cards;
detérioration élevé 1
detérioration faible 11
stable élevé 13
stable faible 53
légère_amélioration élevé 16
légère_amélioration faible 42
amélioration_modérée élevé 15
amélioration_modérée faible 27
amélioration_marquée élevé 7
amélioration_marquée faible 11
;
run;
```

10. Cette statistique est basée sur le coefficient de la régression pondérée des proportions observées sur les scores. Il faut naturellement adapter l'expression du texte lorsque la table est $2 \times C$.

Frequency Percent Row Pct Col Pct	Table of evolution by dosage		
	dosage		
evolution	élevé	faible	Total
détérioration	1 0.51 8.33 1.92	11 5.61 91.67 7.64	12 6.12
stable	13 6.63 19.70 25.00	53 27.04 80.30 36.81	66 33.67
légère_amélioration	16 8.16 27.59 30.77	42 21.43 72.41 29.17	58 29.59
amélioration_modérée	15 7.65 35.71 28.85	27 13.78 64.29 18.75	42 21.43
amélioration_marquée	7 3.57 38.89 13.46	11 5.61 61.11 7.64	18 9.18
Total	52 26.53	144 73.47	196 100.00

TABLE 5.12 – Tableau croisé sur les données d'exemple de Cochran

Statistics for Table of evolution by dosage

Cochran-Armitage Trend Test	
Statistic (Z)	2.5818
Asymptotic Test	
One-sided Pr > Z	0.0049
Two-sided Pr > |Z|	0.0098
Exact Test	
One-sided Pr >= Z	0.0063
Two-sided Pr >= |Z|	0.0106

TABLE 5.13 – Résultats du test de Cochran-Armitage

Partant de là, l'exécution du programme suivant renvoie en particulier la table 5.12. Si on regarde les proportions afférentes aux lignes (*Row Pct*), on peut voir une augmentation de la probabilité d'avoir reçu une dose élevée lorsque l'on descend dans la table, i.e., lorsque l'état du malade s'améliore [11]. L'hypothèse nulle est que pour un dosage donné, la probabilité d'être dans l'une quelconque des 5 modalités ne varie pas. L'alternative avec le test de Cochran-Armitage est que la probabilité d'avoir reçu une dose élevée de traitement augmente avec l'amélioration de l'état du patient.

```
proc freq data=cochran order=data;
tables evolution*dosage / trend;
weight effectif;
exact trend / maxtime=30;
run;
```

La table 5.13 est générée par l'option `trend` de la commande `tables` et la commande `exact trend`. Les deux probabilités critiques, asymptotique et exacte, permettent de conclure au rejet de H0 et à la dépendance (linéaire)

11. Par exemple, la probabilité estimée d'avoir reçu une dose élevée lorsque l'état s'est détérioré est de 8.33%, elle passe à 38.89% lorsque l'amélioration est marquée.

Monte Carlo Estimates for the Exact Test	
One-sided Pr >= Z	
Estimate	0.0067
99% Lower Conf Limit	0.0046
99% Upper Conf Limit	0.0088
Two-sided Pr >= |Z|	
Estimate	0.0120
99% Lower Conf Limit	0.0092
99% Upper Conf Limit	0.0148
Number of Samples	10000
Initial Seed	824775001

TABLE 5.14 – Probabilités critiques du test de Cochran-Armitage obtenues par rééchantillonnage

entre le dosage et l'évolution de l'état du patient, celle-ci étant mesurée sur une échelle ordinale croissante ce qui naturellement plaide en faveur des dosages élevés.

Pour clore ce point, nous avons également réclamé l'estimation des probabilités critiques par rééchantillonnage. La commande `exact` précédente est remplacée par

```
exact trend / mc;
```

La table 5.14 est alors affichée. Les échantillons bootstrappés sont constitués en maintenant fixes les marges de la table [12]. Les probabilités critiques sont estimées par la proportion d'échantillons bootstrappés qui donnent une statistique T plus extrême que celle calculée sur l'échantillon initial. Dans le cas présent, on voit que cette méthode favorise également l'hypothèse de la dépendance des deux variables.

Le test de Cochran-Armitage via la proc Logistic

L'équation (5.6) rappelle que le test de Cochran-Armitage impose une dépendance linéaire entre la probabilité d'un événement et un score. Dans cette équation, Pr[réponse=1] $= \beta_0 + \beta_1 s_i$, l'absence de dépendance revient à poser $\beta_1 = 0$. Or on sait que l'estimation de ce type d'équation peut se faire en ajustant un modèle logistique : on doit donc pouvoir réaliser le test de cette hypothèse nulle en passant par la proc logistic. Comme on travaille sous l'hypothèse nulle pour obtenir la statistique T et la z-transform associée dans la proc freq, on peut penser que dans la proc logistic, le test du score sur H0 : $\beta_1 = 0$ devrait fournir un test équivalent et conduire à la même probabilité critique. Ceci peut se vérifier en exécutant le programme suivant sur les mêmes données que précédemment :

```
data cochran2;
  set cochran;
  if dosage = "élevé" then reponse=1;
```

12. Il est naturellement possible de préciser un `seed` ainsi que le nombre d'échantillons bootstrappés.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.6514	1	0.0099
Score	6.6657	1	0.0098
Wald	6.4651	1	0.0110

TABLE 5.15 – Test de Cochran-Armitage via le test du score dans la proc logistic

```
    else reponse=0;
if evolution ="detérioration" then score=1;
  else if evolution ="stable" then score=2;
  else if evolution ="légère_amélioration" then score=3;
  else if evolution ="amélioration_modérée" then score=4;
  else if evolution ="amélioration_marquée" then score=5;
run;
proc logistic data=cochran2 descending;
  model reponse = score;
  freq effectif;
run;
```

L'étape data crée la variable réponse (1 si le malade a reçu une dose élevée, 0 sinon) et les scores identiques à ceux qu'utilisait la proc freq. La proc logistic renvoie la table 5.15 et on constate bien que la probabilité critique du test du score, 0.0098, est bien la probabilité critique bilatérale que renvoyait la proc freq pour le test de Cochran-Armitage dans la table 5.13 [13].

13. Pour plus de détails, on peut consulter l'article "Cochran-Armitage Trend Test Using SAS" de Hui Liu disponible sur le net. Il montre comment le test de Cochran-Armitage peut être indifféremment réalisé avec la proc freq, la proc logistic et la proc multtest avec pour chacune, calcul des probabilités critiques exactes et/ou estimées par rééchantillonnage.

Chapitre 6

Les mesures d'association

Le test du Khi2 qui est utile pour réaliser le test d'une hypothèse de non association entre deux variables dont les observations ont été regroupées en classes n'est toutefois pas exempt de défauts. En particulier sa valeur est dépendante d'une part de la taille du tableau sur lequel la statistique est évaluée, et d'autre part, de l'effectif total n. Ainsi, pour de grandes valeurs de n, on est conduit quasi systématiquement à rejeter l'hypothèse nulle. Dépendant aussi de la structure du tableau, il ne permet pas de comparer les Khi2 associés à des tableaux dont les nombres de lignes et de colonnes sont différents. Par ailleurs il ne donne pas de mesure de l'intensité de la dépendance en cas de rejet. Plusieurs statistiques ont été proposées afin de pallier ces inconvénients, les plus utilisées sont le coefficient phi de Pearson et surtout le V de Cramer, qui toutes deux sont dérivées du Khi-deux de Pearson. Ces mesures permettent en outre de comparer les χ^2 tirés de différents tableaux de contingence. Pour les obtenir, il suffit de mettre l'option `/ Chisq` dans la commande `Tables` de la proc Freq.

6.1 le coefficient phi

Il ne s'applique qu'à des tableaux 2×2 [1]. C'est simplement la racine de la statistique de Pearson divisée par l'effectif total. On peut l'interpréter comme un coefficient de corrélation entre deux variables dichotomiques. En particulier, si on l'élève au carré, il donne une mesure de la variance commune aux deux variables :

$$\phi^2 = \frac{Q_P}{n} \tag{6.1}$$

Dans le tableau (2x2) suivant :

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

le coefficient *phi* se calcule encore comme :

$$\phi = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{(n_{11} + n_{12})(n_{21} + n_{22})(n_{11} + n_{21})(n_{12} + n_{22})}} \tag{6.2}$$

1. Il peut dépasser l'unité pour des tableaux de plus grandes dimensions

Ainsi, ϕ est positif si l'association dominante correspond à la diagonale principale, *i.e.* si la dépendance est positive entre les deux variables, négatif si elle relève de l'anti-diagonale. Il vaut zéro en cas d'absence de dépendance et est égal à 1 ou -1 lorsque leur dépendance est parfaite.

6.2 le coefficient V de Cramer

Il peut être employé pour mesurer l'intensité des associations entre deux variables nominales ou ordinales dans les tableaux de tailles supérieures à 2×2. De ce fait il est plus général que le coefficient ϕ qui vient d'être présenté et se définit comme la racine carrée du khi-deux de Pearson rapportée au χ^2 maximal théorique que pourrait fournir le tableau de contingence étudié. Dans une table $R \times C$, ce χ^2 de Pearson maximal est donné par $n\,[\min(R-1, C-1)]$. En effet, on a

$$\begin{aligned}
Q_P &= \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{(o_{r,c} - e_{r,c})^2}{e_{r,c}} \\
&= \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{o_{r,c}^2 + e_{r,c}^2 - 2o_{r,c}e_{r,c}}{e_{r,c}} \\
&= \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{o_{r,c}^2}{e_{r,c}} + e_{r,c} - 2o_{r,c} \\
&= \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{o_{r,c}^2}{e_{r,c}} + \sum_{r=1}^{R}\sum_{c=1}^{C} e_{r,c} - 2\sum_{r=1}^{R}\sum_{c=1}^{C} o_{r,c} \qquad (6.3) \\
&= \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{o_{r,c}^2}{e_{r,c}} + n - 2n \\
&= \sum_{r=1}^{R}\sum_{c=1}^{C} \left[\frac{o_{r,c}^2}{e_{r,c}} - n\right] \\
&= n\sum_{r=1}^{R}\sum_{c=1}^{C} \left[\frac{o_{r,c}^2}{o_r o_c} - 1\right] \text{ car sous } H0, e_{rc} = \frac{o_r}{n}o_c = \frac{o_c}{n}o_r
\end{aligned}$$

comme la somme des effectifs sur une ligne, o_r, ou sur une colonne, o_c, est obligatoirement supérieure à l'effectif d'une des cellules de la ligne ou de la colonne concernée, on a $\frac{o_{r,c}}{o_c} \leq 1$ et $\frac{o_{r,c}}{o_r} \leq 1$. Si on utilise la deuxième inégalité, il vient :

$$\sum_{r=1}^{R}\sum_{c=1}^{C} \left[\frac{o_{r,c}^2}{o_r o_c}\right] \leq \sum_{r=1}^{R}\sum_{c=1}^{C} \left[\frac{o_{r,c}}{o_c}\right] = R, \qquad (6.4)$$

avec la première on aurait :

$$\sum_{r=1}^{R}\sum_{c=1}^{C} \left[\frac{o_{r,c}^2}{o_r o_c}\right] \leq \sum_{r=1}^{R}\sum_{c=1}^{C} \left[\frac{o_{r,c}}{o_r}\right] = C \qquad (6.5)$$

Finalement,

$$Q_P = n \sum_{r=1}^{R} \sum_{c=1}^{C} \left[\frac{o_{r,c}^2}{o_r o_c} - 1 \right] \leq n(C-1), \text{et} \tag{6.6}$$

$$Q_P = n \sum_{r=1}^{R} \sum_{c=1}^{C} \left[\frac{o_{r,c}^2}{o_r o_c} - 1 \right] \leq n(R-1) \tag{6.7}$$

L'expression de cette statistique est donc :

$$V = \sqrt{\frac{Q_P}{n\,[\min(R-1, C-1)]}} \tag{6.8}$$

Elle vaut 1 en cas de parfaite dépendance et est égale par construction au coefficient ϕ pour une table 2×2. Sa valeur ne dépend plus de l'effectif total n et elle peut donc être utilisée pour comparer l'intensité de la dépendance entre variables différentes.

Les règles d'interprétation usuelles sont les suivantes :
. $0 \geq V \geq 0.1$: très faible, voir absence de dépendance,
. $0.1 < V \geq 0.3$: dépendance modérée,
. $0.3 < V \geq 0.5$: dépendance assez forte,
. $0.5 < V$: dépendance forte.

6.3 Le coefficient de corrélation de Pearson

Il s'agit de la mesure d'association la plus populaire et vous savez que sur deux variables x et y, il est définit par

$$\rho = \text{cor(x,y)} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{E\left[(x - E[x])\,(y - E[y])\right]}{\sqrt{E[(x - E(x))^2]}\,\sqrt{E[(y - E(y))^2]}} \tag{6.9}$$

et estimé selon :

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6.10}$$

où $\bar{x}$ et $\bar{y}$ sont les moyennes des deux échantillons.
Les deux variables doivent posséder au moins une échelle d'intervalle. Vous savez aussi qu'il varie entre -1 et +1 et n'est utile que pour mesurer l'intensité de relations uniquement linéaires et donc parfaitement adapté pour révéler des dépendances entre gaussiennes. En revanche, il ne l'est plus lorsque les dépendances ne sont plus linéaires.
Naturellement, en tant qu'estimateur du maximum de vraisemblance, $\hat{\rho}$ est asymptotiquement gaussien, mais à distance finie, il possède une distribution asymétrique avec un sens d'asymétrie dépendant du signe de ρ. Fisher a alors proposé une transformation visant à obtenir une variable dont la distribution se rapproche de celle d'une gaussienne lorsque n augmente :

$$z = .5 \log\left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}}\right) \tag{6.11}$$

dont la variance est indépendante de ρ : $var(z) = (n-3)^{-1}$. Lorsque l'option `fisher` est présente dans l'appel de la `proc corr`, celle-ci construit un intervalle de confiance sur la corrélation en passant par cette transformation avec calcul des bornes sur z via la distribution normale puis retour sur $\hat{\rho}$ en appliquant l'inverse de la transformation de fisher sur ces bornes. Elle évalue également les probabilités critiques associées au test de $H0 : \rho = \rho_0$ avec ou sans la transformation de fisher selon la présence ou l'absence de cette option dans son appel. L'alternative $H1$ est gérée par la syntaxe `type=lower | upper | twosided` et la valeur de ρ_0 par une option `rho0`= associée à `fisher`. Ainsi,

. `proc corr data=XXX fisher(alpha=0.10 rho0=0.2 type=lower);`
var x y ;
va donner la borne inférieure d'un intervalle de confiance à 90% sur ρ et une évaluation de la probabilité critique de $H0 : \rho = 0.2$ associée à $H1 : \rho <= 0.2$ avec l'emploi de la transformation de fisher.

. `proc corr data=XXX;`
var x y ;
renvoie la valeur de l'estimateur de Pearson $\hat{\rho}$ et la probabilité critique asymptotique du test $H0 : \rho = 0.0$ versus $H1 : \rho \neq 0$. Lorsqu'aucune autre mesure de corrélation n'est demandée, celle de Pearson est calculée par défaut. Lorsqu'une autre mesure est explicitement réclamée, il est nécessaire de faire apparaître le mot clef `pearson` dans l'appel de la proc pour l'obtenir.

6.4 Le coefficient de corrélation de Spearman

On considère deux variables mesurées au moins sur une échelle ordinale. Si on note respectivement r_x et r_y les variables contenant les rangs des observations de x et y, alors le coefficient de corrélation de Spearman est donné par :

$$\hat{\rho}_S = \frac{\sum_{i=1}^{n}(rx_i - \bar{rx})(ry_i - \bar{ry})}{\sqrt{\sum_{i=1}^{n}(rx_i - \bar{rx})^2}\sqrt{\sum_{i=1}^{n}(ry_i - \bar{ry})^2}} \tag{6.12}$$

Ce n'est donc rien d'autre que la corrélation de Pearson évaluée non pas sur les observations elles-mêmes mais sur les rangs des observations des deux variables [2]. Il fait référence aux notions de paires discordantes ou concordantes, *i.e.* soit les deux paires d'observations (x_i, y_i) et (x_j, y_j), on dira que ces paires sont

— concordantes si $x_i < x_j$ et $y_i < y_j$ ou $x_i > x_j$ et $y_i > y_j$
— discordantes si $x_i < x_j$ et $y_i > y_j$ ou $x_i > x_j$ et $y_i < y_j$

C'est en effet un estimateur de la quantité ρ_S définie par

$$\rho_S = 12E\Big[I(x_j < x_i)I(y_k < y_i)\Big] - 3 \text{ pour tous les } 1 \leq i < j < k \leq n$$

expression dans laquelle $I()$ est la fonction indicatrice usuelle. On peut montrer que sur les deux variables x et y, $\rho_S = 1$ lorsque toutes leurs paires d'observations sont concordantes, et ρ_S=-1 si elles sont toutes discordantes.

Son intérêt vient du fait que des paires concordantes ou discordantes peuvent être créées par des dépendances linéaires ou non entre les variables. Par exemple

2. En cas d'égalité de rangs, `proc cor` remplace ceux-ci par leur rang moyen.

si y est une transformée monotone mais non linéaire de x alors cette dépendance sera parfaitement détectée par la corrélation de Spearman alors qu'elle peut ne pas l'être par la corrélation de Pearson. On préférera donc Spearman lorsque l'on pense que les dépendances ne sont pas linéaires ou/et que les variables ne sont pas gaussiennes. Aucune hypothèse paramétrique n'étant faite sur x et y, $\hat{\rho}_S$ est un estimateur non paramétrique de ρ_S. Sa lecture est identique à celle de $\hat{\rho}$: il varie entre -1 et +1 selon le sens et l'intensité de la dépendance et vaut 0 en cas d'indépendance.
`proc corr` évalue la probabilité critique de $H0 : \rho_S = 0.0$ versus $H1 : \rho_S \neq 0$ en considérant que la transformée $t = (n-2)^{1/2}\left(\frac{\hat{\rho}_S^2}{1-\hat{\rho}_S^2}\right)^{1/2}$ est la réalisation d'une student à (n-2) degrés de liberté.

6.5 Le coefficient de corrélation de Kendall

Ce coefficient, nommé tau-b, fait explicitement référence aux paires d'observations de x et y discordantes et concordantes. En l'absence de rangs égaux dans les deux variables, il s'agit simplement de l'écart entre le nombre de paires concordantes et discordantes rapporté au nombre total de paires d'observations, soit :

$$\hat{\tau}_b = \frac{n_c - n_d}{n_t}, \text{ avec } n_t = n(n-1)/2 \tag{6.13}$$

Avec cette expression, on voit aisément que lorsque $n_c = n_t$, *i.e.* $n_d = 0$ alors $\hat{\tau}_b = 1$, et que lorsque toutes les paires sont discordantes alors $\hat{\tau}_b = -1$. Enfin lorsqu'il y a indépendance le nombre de paires concordantes doit être proche du nombre de paires discordantes, *i.e.* $\hat{\tau}_b =$ doit être proche de zéro. Elle varie donc entre -1 et 1 et sa valeur s'interprète comme les deux corrélations vues précédemment.
Dans la `proc corr`, l'expression précédente est corrigée pour tenir compte des groupes d'observations de même rang au sein des variables x et y. La corrélation de Kendall est évaluée selon

$$\hat{\tau}_b = \frac{\sum_{i<j}\left[sgn(x_i - x_j)sgn(y_i - y_j)\right]}{\sqrt{(n_t - n_x)(n_t - n_y)}}, \tag{6.14}$$

où n_x(resp. n_y) $= \sum_k n_k(n_k - 1)/2$ et n_k est le nombre d'observations égales au sein du $k^{\text{ème}}$ groupe de valeurs égales dans x (resp. dans y) et où $sgn(z)$ est définie par

$$sgn(z) = \begin{cases} 1 \text{ si } z > 0 \\ 0 \text{ si } z = 0 \\ -1 \text{ si } z < 0 \end{cases}$$

Le tau-b de Kendall est un estimateur de la quantité :

$$\tau_b = 2E\left[I(x_i < x_j)I(y_i < y_j)\right] - 1 \text{ pour tous les } 1 \leq i < j \leq n$$

dont on peut aisément voir qu'elle vaut 1 si toutes les paires sont concordantes, -1 lorsque toutes sont discordantes et 0 s'il y a indépendance de x et y. Comme

la corrélation de Spearman, celle de Kendall est une statistique non paramétrique. Étant fondée comme elle sur la concordance ou discordance des paires d'observations, son emploi est recommandé dans le cas de dépendance non linéaire entre les deux variables et/ou sur des variables non gaussiennes. Dans la `proc corr`, le calcul de la probabilité critique associée à $H0 : \tau_b = 0$ est réalisé via une z-transformation sur $\hat{\tau}_b$.

6.5.1 Illustrations des différences entre Pearson, Spearman et Kendall

Dans les exemples considérés maintenant, il existera une dépendance parfaite entre x et y. Il s'agit simplement de montrer la différence de capacité des statistiques $\hat{\rho}$, $\hat{\rho}_S$ et $\hat{\tau}_b$ à détecter cette dépendance et cela en relation avec les remarques qui précèdent et sachant donc que ceci va distinguer la corrélation de Pearson des deux autres.

1. cas d'une dépendance monotone : si on génère les observations de deux variables x et y au moyen du programme suivant :

   ```
   data SIMUL(drop=i);
   call streaminit(123);
   do i = 1 to 15;
   x = rand("Normal");
   y = x**7;
   output;
   end;
   ```

   alors la commande

   proc corr data=simul pearson spearman kendall ;run ;

   renvoie des coefficients de corrélation de Pearson, de Spearman et de Kendall estimés respectivement égaux à 0.630,1. et 1.

2. cas d'une dépendance parfaite non monotone : si on crée x et y selon :

   ```
   data SIMUL(drop=i);
   call streaminit(123);
   do i = 1 to 15;
   x = rand("Normal");
   y = sin(x**7);
   output;
   end;
   ```

   On récupère alors $\hat{\rho} = 0.304$, $\hat{\rho}_S = 0.621$ et $\hat{\tau}_b = 0.714$. Par ailleurs au test de $H0 : \rho = 0$ est associé une probabilité critique de 27%, alors qu'elle est de 1.3% sur $Ho : \rho_S = 0$ et de .002% sur $H0 : \tau_b = 0$. Aux seuils de risque usuels, la nullité de la corrélation ne serait donc pas rejetée avec Pearson alors qu'elle l'est avec Spearman et Kendall.