Introduction à la décision statistique
Fiches aide-mémoire

Francis Colin, Catherine Collet, Fabrice Dessaint,
Vincent Ginot, Christelle Hennequet, Kiên Kiêu,
François Laurens, Annick Moisan, Pierre Montpied,
Catherine Ravel, Brigitte Schaeffer, Pierre Wavresky

Novembre 1997

Table des matières

Préface

La première édition de ces fiches a été rédigée par

B. Aurousseau	C. Chabanet	G. Fouilloux
P. Gasqui	H. Goyeau	F. Lefèvre
M. Lefort-Buson	J.-P. Ley	J.-M. Membré
A. Moisan	J.-C. Pierrat	F.-X. Oury
C. Roux	F. Volaire.	

Dans cette seconde édition, nous avons ajouté des rubriques sur la mise en oeuvre des différents tests décrits sous SAS et S-PLUS. Les données utilisées dans ces nouvelles rubriques reprennent les données traitées en exemple. Tous les tests ne sont pas couverts : en effet, pour certains d'entre eux, il n'y a pas de procédure SAS ou de fonction S-PLUS leur correspondant. Enfin, nous avons ajouté trois «fiches» sur les tests de corrélation.

Première partie

Un échantillon isolé, ou deux échantillons appariés

Chapitre 1

Tests sur la position, ou comparaison des positions

1.1 Test t de Student

1.1.1 Définition du problème : hypothèses nulle et alternative

Tester l'égalité de la moyenne μ à une valeur donnée (connue) μ_0, avec un n-échantillon de variables X_i de moyenne μ et d'écart-type σ inconnus :

$H_0 \ : \ \mu = \mu_0$ contre $H_1 \ : \ \mu \neq \mu_0$ (alternative bilatérale), ou bien :
$H'_0 \ : \ \mu \leqslant \mu_0$ contre $H'_1 \ : \ \mu > \mu_0$ (alternative unilatérale), ou :
$H''_0 \ : \ \mu \geqslant \mu_0$ contre $H''_1 \ : \ \mu < \mu_0$ (alternative unilatérale).

Le test de la moyenne d'un échantillon peut se généraliser à la comparaison des moyennes de deux échantillons appariés : il s'agit souvent de deux variables, X_i et Y_i, mesurées sur les mêmes unités expérimentales. On teste alors l'égalité à 0 de la moyenne des différences $(X_i - Y_i)$.

1.1.2 Conditions d'utilisation

Il faut avoir des variables continues à distribution gaussienne. Ce test reste approximativement valide pour d'autres distributions, à condition que la taille de l'échantillon soit suffisante (à partir de 30 environ).

Pour la comparaison des moyennes de deux échantillons appariés, les n couples de variables $(X_i,\ Y_i)$ doivent être indépendants, et les différences $(X_i - Y_i)$ doivent suivre la même distribution gaussienne.

1.1.3 Définition de la statistique, justification intuitive

On définit la statistique t à partir des statistiques $\widehat{\mu} = (\sum_{i=1}^{n} X_i)/n$ et $\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \widehat{\mu})^2/(n-1)$:

$$t = \frac{\widehat{\mu} - \mu_0}{\widehat{\sigma}/\sqrt{n}}$$

Intuitivement, elle mesure l'écart entre la moyenne testée (0 pour la différence entre deux échantillons appariés) et la moyenne observée, en tenant compte, avec l'écart-type mesuré $\widehat{\sigma}$, de la variabilité propre du phénomène.

1.1.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $|t| > c_{\alpha/2}$, ou bien :
On rejette H_0' au profit de H_1' si $t > c_\alpha$, ou :
On rejette H_0'' au profit de H_1'' si $t < -c_\alpha$.
Les valeurs critiques c_α (ou $c_{\alpha/2}$) sont les quantiles $t_{1-\alpha}$ (ou $t_{1-\alpha/2}$) d'une variable t de STUDENT à $n-1$ degrés de liberté (cf. sect. 1.1.8).

1.1.5 Mise en œuvre

Calcul de la statistique
La variable $\widehat{\sigma}$ peut se calculer avec $\widehat{\sigma}^2 = [\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n]/(n-1)$.

On peut également calculer t à partir des différences $D_i = X_i - \mu_0$ (ou $D_i = X_i - Y_i$) : $t = (\sum_{i=1}^{n} D_i/n)/\sqrt{\{\sum_{i=1}^{n}[D_i - (\sum_{i=1}^{n} D_i)/n]^2/n(n-1)\}}$.

Distribution sous H_0
La variable t suit, sous H_0, une loi de STUDENT à $n-1$ degrés de liberté.

1.1.6 Compléments

Pour une distribution non-gaussienne mais symétrique, voir le test des signes et rangs de WILCOXON ; si la symétrie n'est pas assurée non plus, voir le test du signe.

1.1.7 Exemple

(*Tiré de* CONOVER, 1980).

Pour tester l'effet de l'alcool au volant, 20 conducteurs et conductrices ont été échantillonnés. On a mesuré leur temps de réaction (en secondes) à jeun puis après ingestion d'un doux breuvage alcoolisé. Pour chacune des 20 personnes, les temps mesurés avant et après ont été les suivants :

Avant	0,68	0,64	0,68	0,82	0,58	0,80	0,72	0,65	0,84	0,73
Après	0,73	0,62	0,66	0,92	0,68	0,87	0,77	0,70	0,88	0,79
Avant	0,65	0,59	0,78	0,67	0,65	0,76	0,61	0,86	0,74	0,88
Après	0,72	0,60	0,78	0,66	0,68	0,77	0,72	0,86	0,72	0,97

Soit D_i la différence entre les temps de réponse du i-ième conducteur (*après* − *avant*). On teste l'hypothèse nulle : *les D_i ont une moyenne nulle* (ou éventuellement négative, *i. e.* le breuvage testé n'augmente pas le temps de réponse) contre une alternative unilatérale : *les D_i ont une moyenne strictement positive* (on pense bien *a priori* que l'alcool ne va pas diminuer le temps de réponse). Les 20 valeurs des D_i sont les suivantes :

0,05	−0,02	−0,02	0,10	0,10	0,07	0,05	0,05	0,04	0,06
0,07	0,01	0,00	−0,01	0,03	0,01	0,11	0,00	−0,02	0,09

$\widehat{\mu} = \sum D_i/n = 0,038\,5$, $\widehat{\sigma} = 0,043\,2$, $t = 0,038\,5/(0,043\,2/\sqrt{20}) = 3,986$.

Pour $n - 1 = 19$, et un risque $\alpha = 0,05$ (5 %), la table donne la valeur critique $c_\alpha = 1,729$, supérieure à t : on rejette l'hypothèse nulle.

1.1.8 Table de quantiles des distributions t de Student

La table donne les quantiles $t_{1-\alpha}$ d'un t de STUDENT à *ddl* degrés de liberté.

α :	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
ddl = 1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	127,32	318,31	636,62
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,598
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,214	12,924
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850

suite à la page suivante ...

... suite

α :	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,767
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,255	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	0,254	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
120	0,254	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,253	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

1.1.9 Mise en œuvre informatique

S-PLUS Utiliser la fonction `t.test`.

```
> avant<-c(0.68,0.64,0.68,0.82,0.58,0.80,0.72,
           0.65,0.84,0.73,0.65,0.59,0.78,0.67,
           0.65,0.76,0.61,0.86,0.74,0.88)
> apres<-c(0.73,0.62,0.66,0.92,0.68,0.87,0.77,
           0.70,0.88,0.79,0.72,0.60,0.78,0.66,
           0.68,0.77,0.72,0.86,0.72,0.97)
> res<-t.test(apres,avant,alternative="greater",paired=T)
> summary(res)

Test de Student, deux echantillons apparies :

test de l'hypothese nulle :
*H0 : "la moyenne des differences est egale a 0",
contre l'hypothese alternative unilaterale :
*H1 : "la moyenne des differences est superieure a 0".

Conditions d'utilisation :
elles portent sur la difference des deux echantillons :
        les observations sont independantes
        (tirage au sort des unites experimentales),
        la distribution de la variable aleatoire est normale.

------------------------------------------------------------
         Paired t-Test

data:  apres and avant
t = 3.9858, df = 19, p-value = 4e-04
alternative hypothesis: true mean of differences is greater
than 0
```

```
95 percent confidence interval:
 0.02179774         NA
sample estimates:
 mean of x - y
        0.0385

------------------------------------------------------------
Le test est realise au niveau de confiance 95%.
On rejette l'hypothese nulle, le risque d'erreur est egal
a 4e-04.
```

Le risque d'erreur indiqué est le risque de première espèce (coïncide avec la P-variable).

SAS Procédure **means**

```
data apparie;
     input avant apres @@;
     diff=apres-avant;
     cards;
     0.68 0.73     0.64 0.62     0.68 0.66     0.82 0.92
     0.58 0.68     0.80 0.87     0.72 0.77     0.65 0.70
     0.84 0.88     0.73 0.79     0.65 0.72     0.59 0.60
     0.78 0.78     0.67 0.66     0.65 0.68     0.76 0.77
     0.61 0.72     0.86 0.86     0.74 0.72     0.88 0.97
;

proc means n mean stderr t prt;
     var diff;
run;
```

```
    Analysis Variable : DIFF
     N          Mean     Std Error             T  Prob>|T|
    -----------------------------------------------------
    20     0.0385000     0.0096593     3.9857842    0.0008
    -----------------------------------------------------
```

remarque : les résultats donnés correspondent à un test bilatéral, il faut diviser par 2 la p-value pour avoir la valeur du test unilatéral.

1.2 Test des signes et rangs de Wilcoxon

1.2.1 Définition du problème : hypothèses nulle et alternative

Tester l'égalité de la médiane m à une valeur donnée m_0 (connue), avec un n-échantillon de variables X_i de médiane m inconnue :

	H_0 :	$m = m_0$	contre	H_1 :	$m \neq m_0$	(alternative bilatérale)
ou bien	H_0' :	$m \leq m_0$	contre	H_1' :	$m > m_0$	(alternative unilatérale)
ou bien	H_0'' :	$m \geq m_0$	contre	H_1'' :	$m < m_0$	(alternative unilatérale).

Le test sur la médiane d'un échantillon peut se généraliser à la comparaison des médianes de deux échantillons appariés (il s'agit souvent de deux variables, X_i et Y_i, mesurées sur les mêmes unités expérimentales). On teste alors l'égalité à 0 de la médiane des différences $(X_i - Y_i)$.

1.2.2 Conditions d'utilisation

Ce test ne peut pas s'utiliser avec de très petits échantillons ($n < 6$), car on ne peut alors définir de région de rejet dont la probabilité (risque de première espèce) soit inférieure à 5 %.

Il faut avoir des variables continues ou discrètes en catégories ordonnées, dont on peut supposer *a priori* que la distribution est symétrique.

Pour la comparaison des médianes de deux échantillons appariés, les n couples de variables $(X_i,\ Y_i)$ doivent être indépendants, et les différences $(X_i - Y_i)$ doivent suivre la même distribution symétrique (ce que l'on peut obtenir en particulier si les X_i et Y_i suivent la même distribution —à un décalage de position près sous l'hypothèse alternative H_1— même si celle-ci n'est pas symétrique).

1.2.3 Définition de la statistique, justification intuitive

Soit n le nombre d'observations X_i différentes de m_0 (les éventuelles observations identiques à m_0 ne sont pas prises en compte, et on appelle n la taille de l'échantillon réduit).

À partir des n variables X_i initiales, on peut définir $n(n+1)/2$ variables $Z_{ij} = (X_i + X_j)/2$ (pour $1 \leqslant i \leqslant j \leqslant n$), qui sont les moyennes deux à deux des variables initiales (on les appelle parfois « moyennes de WALSH »).

On montre que si la distribution des X_i est symétrique, alors la médiane de la distribution de Z_{ij} est la même que celle de X_i. On fera donc un test d'égalité de la médiane des Z_{ij} à la valeur m_0.

Soit W^+ le nombre de valeurs de variables Z_{ij} supérieures à m_0, et
W^- le nombre de valeurs de variables Z_{ij} inférieures à m_0.

On a $W^+ + W^- = n(n+1)/2$ (nombre total de valeurs comparées à m_0). Sous l'hypothèse nulle, W^+ et W^- ont des valeurs voisines, et donc proches de $n(n+1)/4$, tandis que sous l'hypothèse alternative, l'une de ces deux statistiques (et donc aussi leur minimum) est plus proche de 0.

Remarque : les statistiques de WILCOXON W^+ et W^- sont construites de manière analogue aux statistiques S^+ et S^- du test du signe, mais elles ne suivent pas la même loi (les tables sont donc différentes) car les Z_{ij}, contrairement aux variables X_i de départ, ne sont pas indépendantes (par exemple, Z_{ij} et Z_{ik} dépendent toutes deux de X_i).

1.2.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $\min(W^+, W^-) \leqslant c_{\alpha/2}$, ou bien :
On rejette H'_0 au profit de H'_1 si $W^- \leqslant c_\alpha$, ou :
On rejette H''_0 au profit de H''_1 si $W^+ \leqslant c_\alpha$.

1.2.5 Mise en œuvre

Calcul de la statistique Il y a deux façons de calculer W^+ et W^-:

- On peut calculer directement les $n(n+1)/2$ valeurs des moyennes Z_{ij} (ne pas oublier d'inclure les valeurs initiales $X_i = (X_i + X_i)/2 = Z_{ii}$) et les comparer à m_0. Il est conseillé d'utiliser un tableau comme ci-dessous, dont on ne remplit que le triangle inférieur (ou supérieur) :

Valeurs des X_i *ordonnées*	X_1	X_2	X_3	$\dots$	X_n
X_1	X_1				
X_2	$\frac{X_1+X_2}{2}$	X_2			
X_3	$\frac{X_1+X_3}{2}$	$\frac{X_2+X_3}{2}$	X_3		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
X_n	$\frac{X_1+X_n}{2}$	$\frac{X_2+X_n}{2}$	$\frac{X_3+X_n}{2}$	$\dots$	X_n

- Dès que n n'est pas très petit ($n \geqslant 10$), la méthode suivante est plus pratique : on calcule les n différences $D_i = X_i - m_0$; certaines sont positives, d'autres négatives, on range leurs valeurs absolues $|D_i|$ de 1 à n. Si plusieurs différences ont la même valeur absolue (que les X_i concernés soient ex æquo ou qu'ils soient à égale distance de part et d'autre de m_0), on leur attribue à toutes un même rang moyen (moyenne des rangs qui leur seraient attribuées après leur avoir appliqué de petites perturbations infinitésimales indépendantes par exemple). On obtient alors W^+ en faisant la somme des rangs associés aux différences positives, et W^- en faisant la somme des rangs associés aux différences négatives (la somme des deux $W^+ + W^-$ vaut bien toujours $n(n+1)/2$, qui est ici la somme de tous les rangs de 1 à n).

Distribution sous H_0 Comme nous l'avons remarqué à la fin de la section 1.2.3, la distribution de W^+ (ou de W^-) n'est pas binomiale. On la calcule en supposant équiprobables toutes les 2^n façons d'associer des signes aux rangs des différences D_i (et on note le résultat dans une table, cf. sect. 1.2.8, pour ne pas avoir à recommencer à chaque fois)...

Dans le cas de grands échantillons ($n > 20$) et/ou s'il y a de nombreux ex æquo dans l'échantillon des X_i, on utilise plutôt la statistique $Z = \sum_{i=1}^n R_i / \sqrt{\sum_{i=1}^n R_i^2}$, où R_i est le rang de la valeur absolue $|D_i|$ affecté

du signe de la différence D_i. Les valeurs critiques sont lues dans la table de la loi de GAUSS $\mathcal{N}(0,1)$ (ou sur la ligne « ∞ » de la table de la loi de STUDENT, cf. sect. 1.1.8, p. 5).

1.2.6 Compléments

Si la distribution des X_i (ou des D_i) est gaussienne, alors le test t de STUDENT est plus puissant. Si cette distribution n'est pas symétrique, voir le test du signe.

Une façon plus expéditive de traiter le cas des grands échantillons et/ou d'ex æquo nombreux consiste à appliquer la procédure du test t de STUDENT aux rangs R_i (à la place des différences D_i comme il est fait dans le test t).

Il existe des variantes du test de WILCOXON, fondées sur la statistique $|W^+ - W^-|$, qui ne suit pas la même loi sous H_0 (cf. SPRENT, 1992).

1.2.7 Exemple

(*Tiré de* SPRENT, 1992).

Sur un échantillon de 15 parcelles de surface 1 m^2, le comptage du nombre de plantes malades a donné les valeurs suivantes :
21 17 43 81 32 102 117 43 39 11 67 23 142 7 44
On voudrait savoir si on retrouve bien une médiane de 50 comme dans les études antérieures. On suppose que la distribution est symétrique.

On teste l'hypothèse nulle « *la médiane vaut 50* » contre une alternative bilatérale. Ici, $n = 15$ (aucune valeur n'est égale à 50). On peut calculer les 15 différences $D_i = X_i - 50$, respectivement :
-29 -33 -7 31 -18 52 67 -7 -11 -39 17 -27 92 -43 -6

Les rangs attribués aux valeurs absolues sont respectivement :
8 10 2,5 9 6 13 14 2,5 4 11 5 7 15 12 1,
et donc $W^+ = 9 + 13 + 14 + 5 + 15 = 56$, et

$$W^- = 8 + 10 + 2,5 + 6 + 2,5 + 4 + 11 + 7 + 12 + 1 = 64$$

(on vérifie bien que $56 + 64 = 120 = (15 \times 16)/2$).

Pour $n = 15$ et un risque $\alpha = 0,05$ (5 %), la table nous donne $c_{\alpha/2} = 26$. Comme W^+ et W^- sont toutes les deux supérieures à la valeur critique, on ne peut pas rejeter l'hypothèse nulle.

1.2.8 Table de valeurs critiques du test des signes et rangs

Pour n et α donnés, la table donne la valeur critique c_α (ou $c_{\alpha/2}$).

α ou $\alpha/2 =$	0,005	0,01	0,025	0,05	0,10	0,20	0,30	0,40	0,50
$n = 4$	0	0	0	0	1	3	3	4	5
5	0	0	0	1	3	4	5	6	7,5
6	0	0	1	3	4	6	8	9	10,5
7	0	1	3	4	6	9	11	12	14
8	1	2	4	6	9	12	14	16	18
9	2	4	6	9	11	15	18	20	22,5
10	4	6	9	11	15	19	22	25	27,5
11	6	8	11	14	18	23	27	30	33
12	8	10	14	18	22	28	32	36	39
13	10	13	18	22	27	33	38	42	45,5
14	13	16	22	26	32	39	44	48	52,5
15	16	20	26	31	37	45	51	55	60
16	20	24	30	36	43	51	58	63	68
17	24	28	35	42	49	58	65	71	76,5
18	28	33	41	48	56	66	73	80	85,5
19	33	38	47	54	63	74	82	89	95
20	38	44	53	61	70	83	91	98	105

1.2.9 Mise en œuvre informatique

S-PLUS Utiliser la fonction `wilcox.test`.

```
> ill.plant<-c(21,17,43,81,32,102,117,43,39,11,
               67,23,142,7,44)
> res<-wilcox.test(ill.plant,alternative="two.sided",mu=50)
Warning messages:
  cannot compute exact p-value with ties in: wil.
        sign.rank(dff, alternative, exact,
        correct)
> summary(res)

Test de Wilcoxon, deux echantillons apparies :

test de l'hypothese nulle :
*H0 : "la mediane des differences est egale a 50",
contre l'hypothese alternative bilaterale :
*H1 : "la mediane des differences est differente de 50".

Conditions d'utilisation :
elles portent sur la difference des deux echantillons :
        les observations sont independantes
        (tirage au sort des unites experimentales),
        la distribution de la variable aleatoire est
        symetrique.

------------------------------------------------------------
         Wilcoxon signed-rank test

data:  ill.plant
```

```
signed-rank normal statistic with correction Z = -0.1988,
p-value = 0.8424
alternative hypothesis: true mu is not equal to 50

------------------------------------------------------------
Le test est realise au niveau de confiance 95%.
On ne rejette pas l'hypothese nulle, on accepte l'egalite des
medianes.
```

L'avertissement signale qu'on a des ex-æquo parmi les $|D_i - 50|$. Ceux-ci sont dûs au fait que le nombre de plantes malades par parcelles n'est pas une variable continue. La fonction `wilcox.test` calcule dans ce cas une statistique de test modifiée qui suit approximativement une loi normale centrée réduite. Pour plus de détails sur la correction voir SPRENT, 1992 ou LEHMANN, *Nonparametrics: Statistical Methods Based On Ranks*, Holden and Day, San Francisco, 1975.

SAS Utiliser la procédure **UNIVARIATE**

```
data ill.plant;
     input ill @@;
     diff = ill-50;
     cards;
21 17 43 81 32 102 117 43 39 11 67 23 142 7 44
;

proc univariate;
     var diff;
run;
```

```
                          Univariate Procedure

Variable=DIFF

                               Moments

N                15  Sum Wgts         15
Mean            2.6  Sum              39
Std Dev     40.89499  Variance     1672.4
Skewness    1.018246  Kurtosis   0.106836
USS            23515  CSS         23413.6
CV          1572.884  Std Mean   10.55904
T:Mean=0    0.246234  Pr>|T|       0.8091
Num ^= 0          15  Num > 0           5
M(Sign)         -2.5  Pr>=|M|      0.3018
Sgn Rank          -4  Pr>=|S|      0.8361
W:Normal    0.889299  Pr<W         0.0661
```

La statistique calculée par SAS est :

$$\texttt{Sgn Rank} = S^+ - \frac{n^*(n^*+1)}{4}$$

avec S^+ la somme des rangs des valeurs supérieures à la médiane. La probabilité associée à la statistique est exacte pour $n^* < 20$. Pour $n^* > 20$, on utilise une transformation de la statistique `Sgn Rank` qui suit une distribution t de *Student* avec $n^* - 1$ degrés de liberté.

1.3 Test du signe

1.3.1 Définition du problème: hypothèses nulle et alternative

Tester l'égalité de la médiane m à une valeur donnée m_0 (connue), avec un n-échantillon de variables X_i de médiane m inconnue :

$H_0 \ : \ m = m_0$ contre $H_1 \ : \ m \neq m_0$ (alternative bilatérale), ou bien :

$H'_0 \ : \ m \leqslant m_0$ contre $H'_1 \ : \ m > m_0$ (alternative unilatérale), ou :

$H''_0 \ : \ m \geqslant m_0$ contre $H''_1 \ : \ m < m_0$ (alternative unilatérale).

Le test sur la médiane d'un échantillon peut s'adapter à la comparaison des médianes de deux échantillons appariés (il s'agit souvent des valeurs X_i et Y_i d'une variable, mesurées sur les mêmes unités expérimentales dans des conditions différentes). On teste alors la nullité de la médiane des $(X_i - Y_i)$.

1.3.2 Conditions d'utilisation

Ce test ne peut pas s'utiliser avec de très petits échantillons ($n < 6$), car on ne peut alors définir de région de rejet dont la probabilité (risque de première espèce) soit inférieure à 5 %.

Il faut avoir des variables quantitatives ou qualitatives ordonnées ; mais il n'est pas indispensable de connaître l'ensemble des valeurs exactes : dans le cas d'un échantillon isolé par exemple, il suffit de pouvoir déterminer combien d'observations sont supérieures à m_0 (et combien sont égales).

Pour la comparaison des médianes de deux échantillons appariés, les n couples de variables $(X_i,\ Y_i)$ doivent être indépendants, et les différences $(X_i - Y_i)$ doivent suivre la même distribution.

1.3.3 Définition de la statistique, justification intuitive

Soit n le nombre d'observations X_i différentes de m_0 (les éventuelles observations identiques à m_0 ne sont pas prises en compte, et on appelle n la taille de l'échantillon réduit). Soit S^+ le nombre d'observations supérieures à m_0, et S^- le nombre d'observations inférieures à m_0.

On a $S^+ + S^- = n$ (nombre total d'observations comparées à m_0). Sous l'hypothèse nulle, S^+ et S^- ont des valeurs voisines, et donc proches de $n/2$, tandis que sous l'hypothèse alternative, l'une de ces deux statistiques (et donc aussi leur minimum) est plus proche de 0.

1.3.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $\min(S^+, S^-) \leqslant c_{\alpha/2}$, ou bien H_0' au profit de H_1' si $S^- \leqslant c_\alpha$, ou H_0'' au profit de H_1'' si $S^+ \leqslant c_\alpha$.

1.3.5 Mise en œuvre

Calcul de la statistique
Il n'est pas nécessaire de connaître l'ensemble des valeurs exactes pour calculer les statistiques S^+ ou S^- : il suffit de savoir combien d'observations (resp. de différences) sont supérieures à m_0 (resp. à 0) et combien égales.

Distribution sous H_0
S^+ et S^- suivent toutes deux sous H_0 la même distribution binomiale de paramètres n et $p = \frac{1}{2}$.

Pour les grands échantillons ($n > 20$), on utilise $Z = \frac{S+1/2-n/2}{(\sqrt{n})/2}$, où l'on prend S égale à $\min(S^+, S^-)$ pour l'alternative bilatérale H_1, à S^- pour l'alternative unilatérale H_1' et à S^+ pour l'alternative unilatérale H_1'' ; le 1/2 au numérateur est la « correction de continuité » destinée à affiner l'approximation de la loi binomiale (discrète) par la loi gaussienne (continue). Les valeurs critiques sont alors lues dans la table de la loi de Gauss $\mathcal{N}(0,1)$ (ou sur la ligne « ∞ » de la table de la loi de Student, sect. 1.1.8, p. 5).

1.3.6 Compléments

Lorsque l'on peut supposer *a priori* que la distribution (des X_i ou des $X_i - Y_i$) est symétrique, alors la médiane et la moyenne théoriques (sous l'hypothèse H_0) coïncident, et le test du signe devient aussi un test sur la moyenne. Mais dans ce cas le test des signes et rangs de Wilcoxon est plus puissant : il tient compte de l'importance de chaque écart à m_0.

Plusieurs tests nonparamétriques sont dérivés du test du signe : test de la valeur d'un quantile (cf. Sprent, 1992) ; test de McNemar (comparaison de proportions ou, ce qui revient au même, des positions de variables binaires — ne pouvant prendre que deux valeurs — dans deux échantillons appariés, cf. sect. 1.4) ; test d'une tendance monotone de Cox et Stuart (cf. sect. 2.2), etc.

1.3.7 Exemple

(*Tiré de* Sprent, 1992).

Une entreprise lance un régime amaigrissant avec le slogan : « Perdez 5 kg en 2 mois ! ». Pour vérifier si cette affirmation est vraie « en moyenne », une association de consommateurs a enregistré le poids perdu par chaque individu d'un échantillon de 16 personnes ayant suivi le traitement :

4 6 3 1 2 5 4 0 3 6 3 1 7 2 5 6

On fait le test du signe avec les hypothèses suivantes :
H_0' : *la médiane du poids perdu est au moins 5 kg* (le slogan est justifié), et

H_1' : *cette publicité tend à être mensongère, la médiane du poids perdu est strictement inférieure à 5 kg* (l'alternative est donc unilatérale).

Procédure du test : il y a quatre valeurs supérieures à 5, donc $S^+ = 4$; et il y a deux valeurs égales à 5, donc $n = 16 - 2 = 14$. Pour $n = 14$, au seuil $\alpha = 0,05$ ($p = 0,029$ dans la table), on a $c_\alpha = 3$. Ici, $S^+ > c_\alpha$ donc on ne peut pas rejeter l'hypothèse nulle.

1.3.8 Table de probabilités binomiales cumulées

Chercher, dans la ligne correspondant au n donné, la valeur de p juste inférieure au seuil choisi α (ou $\alpha/2$) ; la valeur critique c_α (ou $c_{\alpha/2}$) se trouve en tête de colonne.

n \ $c =$	0	1	2	3	4	5	6	7	8	9	10
6	0,016	0,109	0,344	0,656	0,891	0,984	1				
7	0,008	0,062	0,227	0,500	0,773	0,938	0,992	1			
8	0,004	0,035	0,144	0,363	0,637	0,856	0,965	0,996	1		
9	0,002	0,020	0,090	0,254	0,500	0,746	0,910	0,980	0,998	1	
10	0,001	0,011	0,055	0,172	0,377	0,623	0,828	0,945	0,989	0,999	1
11	0,001	0,006	0,033	0,113	0,274	0,500	0,726	0,887	0,967	0,994	0,999
12	0,000	0,003	0,019	0,073	0,194	0,387	0,613	0,806	0,927	0,981	0,997
13	0,000	0,002	0,011	0,046	0,133	0,291	0,500	0,710	0,867	0,954	0,989
14	0,000	0,001	0,006	0,029	0,090	0,212	0,395	0,605	0,788	0,910	0,971
15	0,000	0,000	0,004	0,018	0,059	0,151	0,304	0,500	0,696	0,849	0,941
16	0,000	0,000	0,002	0,011	0,038	0,105	0,227	0,402	0,598	0,773	0,895
17	0,000	0,000	0,001	0,006	0,024	0,072	0,166	0,314	0,500	0,686	0,834
18	0,000	0,000	0,001	0,004	0,015	0,048	0,119	0,240	0,407	0,593	0,760
19	0,000	0,000	0,000	0,002	0,010	0,032	0,084	0,180	0,324	0,500	0,676
20	0,000	0,000	0,000	0,001	0,006	0,021	0,058	0,132	0,252	0,412	0,588

n \ $c =$	11	12	13	14	15	16	17	18	19	20
11	1									
12	1,000	1								
13	0,998	1,000	1							
14	0,994	0,999	1,000	1						
15	0,982	0,996	1,000	1,000	1					
16	0,962	0,989	0,998	1,000	1,000	1				
17	0,928	0,976	0,994	0,999	1,000	1,000	1			
18	0,881	0,952	0,985	0,996	0,999	1,000	1,000	1		
19	0,820	0,916	0,968	0,990	0,998	1,000	1,000	1,000	1	
20	0,748	0,868	0,942	0,979	0,994	0,999	1,000	1,000	1,000	1

1.3.9 Mise en œuvre informatique

S-PLUS Il n'y a pas de fonction native dans S-PLUS pour ce test. Toutefois la fonction NESI de la bibliothèque du même nom permet de le mettre en

œuvre.

```
> lost.weight<-c(4,6,3,1,2,5,4,0,3,6,3,1,7,2,5,6)
> library(NESI)
> res<-NESI(lost.weight,alternative="less",mu=5)

"Binom.test(...)"

 Il y a 2 valeurs egales a la valeur de reference

> summary(res)

Test du signe a un echantillon :

test de l'hypothese nulle :
*H0 : "la mediane est egale a 5",
contre l'hypothese alternative unilaterale :
*H1 : "la mediane est inferieure a 5".

Conditions d'utilisation :
      les observations de l'echantillon sont independantes
      (tirage au sort des unites experimentales),
      la distribution de la variable aleatoire est quelconque,
      la taille de l'echantillon doit etre superieure a 6.

------------------------------------------------------------
         Exact binomial test

data:  lost.weight
number of successes = 10, n = 14, p-value = 0.9713
alternative hypothesis: true p is less than 0.5

------------------------------------------------------------
Le test est realise au niveau de confiance 95%.
On ne rejette pas l'hypothese nulle.
```

SAS

```
data apparie;
     input poids @@;
     perte=poids - 5;
     cards;
4 6 3 1 2 5 4 0 3 6 3 1 7 2 5 6
     ;

proc univariate;
     var perte;
run;

                              Univariate Procedure

Variable=PERTE

                                   Moments
```

N	16	Sum Wgts	16
Mean	-1.375	Sum	-22
Std Dev	2.093641	Variance	4.383333
Skewness	-0.07628	Kurtosis	-1.0412
USS	96	CSS	65.75
CV	-152.265	Std Mean	0.52341
T:Mean=0	-2.627	Pr>|T|	0.0190
Num ^= 0	14	Num > 0	4
M(Sign)	-3	Pr>=|M|	0.1796
Sgn Rank	-36	Pr>=|S|	0.0228

La statistique calculée par SAS est :

$$\texttt{M(sign)} = r^+ - \frac{n^*}{2}$$

qui correspond à l'écart entre le nombre de signes «plus» observés r^+ et le nombre attendu $\frac{n^*}{2}$, avec n^* le nombre de valeurs non nulles.

Elle fournit aussi la probabilité cumulée P pour un test bilatéral:

$$\text{Prob} > |\text{M}| = 2 \sum_{i=0}^{min(r^+, n^* - r^+)} C_{n^*}^i \left(\frac{1}{2}\right)^{n^*}$$

Pour avoir la valeur de la p-value du test unilatéral il faut diviser par 2, on obtient donc $Pr >= -M = 0.1796/2 = 0.09$

1.4 Test de McNemar : comparaison de proportions

1.4.1 Définition du problème : hypothèses nulle et alternative

Comparer les positions dans deux échantillons appariés lorsque la variable est binaire (qualitative à seulement deux catégories : succès/échec, oui/non, 0/1, mâle/femelle, etc.). Ce qui revient à comparer les deux proportions (de succès, de oui, de 0, de mâles, etc.) $\Pr\{X_i = 0\}$ et $\Pr\{Y_i = 0\}$, ou encore, en retranchant $\Pr\{X_i = 0, Y_i = 0\}$ de chacune d'elles, à tester :

H_0 : $\forall i, \Pr\{X_i = 0, Y_i = 1\} = \Pr\{X_i = 1, Y_i = 0\}$, contre l'alternative :

H_1 : $\Pr\{X_i = 0, Y_i = 1\} \neq \Pr\{X_i = 1, Y_i = 0\}$ (bilatérale), ou bien :
H_1' : $\Pr\{X_i = 0, Y_i = 1\} > \Pr\{X_i = 1, Y_i = 0\}$ (unilatérale), ou :
H_1'' : $\Pr\{X_i = 0, Y_i = 1\} < \Pr\{X_i = 1, Y_i = 0\}$ (unilatérale).

1.4.2 Conditions d'utilisation

Les données sont un n-échantillon $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, où les X_i et Y_i ne peuvent prendre que deux valeurs, codées arbitrairement 0 et 1. Les données peuvent donc se résumer dans une table de contingence 2×2 :

		Répartition des Y_i	
		$Y_i = 0$	$Y_i = 1$
Répartition des X_i	$X_i = 0$	n_{00}	n_{01}
	$X_i = 1$	n_{10}	n_{11}

1.4.3 Définition de la statistique, justification intuitive

Le test de McNemar est un test du signe appliqué à deux échantillons appariés, où une observation (0, 1) correspond à une différence positive, (1, 0) à une différence négative, et (0, 0) et (1, 1) à des ex æquo (donc ignorés dans la suite de l'analyse).

La statistique de test est simplement le nombre de différences positives (noté n_{01} dans la table de contingence) ou négatives (noté n_{10}) ou le minimum des deux (pour le test bilatéral), l'effectif étant réduit à $n' = n_{01} + n_{10}$. Si les proportions $\Pr\{X_i = 0\}$ et $\Pr\{Y_i = 0\}$ sont égales, les probabilités d'obtenir (0, 1) et (1, 0) le sont aussi : les nombres de différences positives et négatives doivent donc être voisins, et par conséquent proches de $n'/2$.

1.4.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $\min(n_{01}, n_{10}) \leqslant c_{\alpha/2}$, ou bien au profit de H_1' si $n_{10} \leqslant c_\alpha$, ou au profit de H_1'' si $n_{01} \leqslant c_\alpha$.

La région de rejet dépend du niveau de signification α, du test effectué (uni- ou bilatéral), et du nombre n' d'occurrences de (0, 1) ou (1, 0) (mais pas directement de n puisque les $n_{00} + n_{11}$ ex æquo sont ignorés).

1.4.5 Mise en œuvre

Calcul de la statistique

Il ne pose pas de problème une fois les données rangées dans la table 2×2.

Distribution sous H_0

Les statistiques n_{01} et n_{10} suivent sous H_0 la même distribution binomiale de paramètres n' et $p = \frac{1}{2}$. On détermine donc les valeurs critiques c_α ou $c_{\alpha/2}$ dans la table des probabilités binomiales cumulées (cf. sect. 1.3.8).

Pour les grands échantillons ($n' > 20$), on utilise l'approximation gaussienne de la loi binomiale, et la statistique devient (car $n' = n_{01} + n_{10}$) :

$$Z = \frac{n_{01} - n'/2}{(\frac{1}{2}\sqrt{n'})} = \frac{n_{01} - n_{10}}{\sqrt{n_{01} + n_{10}}} \sim \mathcal{N}(0,1)$$

Les valeurs critiques sont alors lues dans la table de la loi de GAUSS $\mathcal{N}(0,1)$ (ou sur la ligne « ∞ » de la table de la loi de STUDENT, cf. sect. 1.1.8, p. 5).

Pour les grands échantillons ($n' > 20$), on peut aussi utiliser la statistique $Z^2 = (n_{01} - n_{10})^2/(n_{01} + n_{10})$, distribuée comme un χ^2 à 1 degré de liberté.

1.4.6 Exemple

(*Tiré de* SPRENT, 1992).

Les membres d'un club d'alpinisme veulent savoir s'il y a une différence de niveau entre deux courses. En consultant le livre d'or du club, ils constatent que les 108 personnes qui ont tenté les deux ascensions se répartissent de la manière suivante :

		Première course	
		Succès	Échec
Seconde course	Succès	73	14
	Échec	9	12

On a donc 14 ascensions tendant à faire penser que la première course est plus difficile, et 9 tendant au contraire. Comme $n' = 14 + 9 = 23$, on peut utiliser l'approximation gaussienne ; la statistique de test vaut alors $Z = (14 - 9)/\sqrt{23} \approx 1,043$. Cette valeur est bien en dessous du seuil de signification ($|Z| > 1,96$ est la région de rejet pour le test bilatéral de niveau 5 %). On ne peut donc pas rejeter l'hypothèse nulle de difficultés égales pour les deux ascensions.

1.4.7 Table

Voir section 1.3.8 (veiller à bien utiliser n' là où la table parle de « n »).

1.4.8 Mise en œuvre informatique

S-PLUS Utiliser la fonction `mcnemar.test`

```
> ascensions<-matrix(c(73,9,14,12),2,2)
> ascensions
     [,1] [,2]
[1,]   73   14
[2,]    9   12
```

```
> res<-mcnemar.test(ascensions)
> res

        McNemar's chi-square test with continuity correction

data:  ascensions
McNemar's chi-square = 0.6957, df = 1, p-value = 0.4042
```

C'est la statistique Z^2 modifiée (qui suit un χ^2 sous H_0) qui est utilisée. La correction de continuité consiste à remplacer $(n_{01} - n_{10})^2$ par $(|n_{01} - n_{10}| - 1)^2$.

Chapitre 2

Tests sur la nature de l'échantillon

2.1 Test du nombre de runs : indépendance d'une suite

2.1.1 Définition du problème : hypothèses nulle et alternative

Déterminer si les n observations d'une suite d'événements aléatoires, de deux sortes possibles, sont indépendantes.

H_0 : les événements aléatoires sont indépendants, contre l'alternative :

H_1 : les événements ne sont pas indépendants (bilatérale), ou bien :

H_1' : les événements ne sont pas indépendants et il y a une tendance au regroupement d'événements semblables (unilatérale), ou :

H_1'' : les événements ne sont pas indépendants et il y a une tendance à l'alternance des deux types d'événements (unilatérale).

2.1.2 Conditions d'utilisation

Les données sont une suite d'observations de variables aléatoires, notées $X_1, X_2, \ldots, X_n$, équidistribuées, et qui sont binaires (oui/non, succès/échec, mâle/femelle, etc.) c'est-à-dire discrètes à valeurs dans un ensemble de deux catégories que l'on peut toujours coder $\{0, 1\}$; par exemple 00 111 0 11 000.

Si les variables ne sont pas binaires mais continues, ou bien discrètes en catégories ordonnées, on s'y ramène en notant 0 une observation strictement inférieure à la médiane de la suite et 1 une observation strictement supérieure à cette médiane, et en ignorant les observations identiques à la médiane (réduisant d'autant la longueur de la suite).

2.1.3 Définition de la statistique, justification intuitive

Dans une suite d'observations on définit un « run » comme une suite de résultats identiques. Par exemple, la suite 00 111 0 11 000 vue plus haut présente 5 runs : 00, 111, 0, 11 et 000.

Le test utilise le nombre total R de runs : « trop » élevé, ce nombre sera révélateur d'une tendance à une alternance des valeurs 0 et 1 plus fréquente que l'hypothèse nulle (indépendance des observations successives) ne permet de le prévoir ; tandis qu'une valeur « trop » faible sera révélatrice d'une autre forme d'écart à l'hypothèse nulle, où des événements de même nature ont tendance à s'agréger (regroupement des 0 et/ou des 1).

2.1.4 Régions de rejet et de non-rejet

Elles sont définies par les valeurs critiques inférieure, R^i, et supérieure, R^s, qui dépendent du niveau α et des nombres de 0 et de 1 de la suite.

On rejette H_0 au profit de H_1 si $R \leqslant R^i_{\alpha/2}$ ou $R \geqslant R^s_{\alpha/2}$, ou bien au profit de H'_1 si $R \leqslant R^i_\alpha$, ou au profit de H''_1 si $R \geqslant R^s_\alpha$.

2.1.5 Mise en œuvre

Calcul de la statistique
Il ne présente aucune difficulté.

Distribution sous H_0
Pour les suites courtes ($\max(n_0, n_1) \leqslant 20$, avec n_0 = nombre de 0 et n_1 = nombre de 1), on détermine les valeurs critiques par lecture dans la table (établie en calculant la distribution à partir de l'énumération de toutes les suites possibles).

Si l'un au moins des deux nombres de 0 ou de 1 dépasse 20, la distribution de R devient approximativement gaussienne et l'on utilise alors la statistique :

$$Z = \frac{R - \mu}{\sigma} \quad \text{avec} \quad \mu = 1 + \frac{2n_0 n_1}{n_0 + n_1} \quad \text{et} \quad \sigma = \frac{\sqrt{2n_0 n_1 (2n_0 n_1 - n_0 - n_1)}}{(n_0 + n_1)\sqrt{(n_0 + n_1 - 1)}},$$

et les valeurs critiques sont lues dans la table de la loi de Gauss $\mathcal{N}(0,1)$ (ou sur la ligne « ∞ » de la table de la loi de Student, cf. sect. 1.1.8, p. 5).

2.1.6 Compléments

D'autres tests de runs ont été proposés :

- la longueur L du run le plus long peut indiquer une tendance au regroupement si elle est « trop » élevée, ou à l'alternance si elle est « trop » faible ; il faut bien sûr une table spécifique pour ce test ;

– dans le cas de données continues, ou de données qualitatives ordonnées, on peut comparer chaque observation non plus à la médiane de la suite, mais à celle qui la précède, en notant le signe de $(X_{i+1} - X_i)$: la suite de n valeurs X_i est alors remplacée par une suite de $n-1$ signes, sur laquelle sont définis des runs « up » (runs de signes $+$) et « down » (runs de signes $-$), dont le nombre total suit une distribution spécifique (car même si les variables de départ sont indépendantes, les signes successifs, eux, ne le sont pas : par exemple la probabilité d'une augmentation, d'un signe $+$, n'est pas la même après une diminution, un signe $-$, ou après une augmentation, un signe $+$) : c'est pourquoi il ne faut bien sûr pas utiliser la même table.

Par contre, le test de runs de WALD-WOLFOWITZ (cf. sect. 3.4), pour comparer deux distributions avec des échantillons indépendants d'une variable continue ou qualitative ordonnée, utilise la même table puisqu'il n'est qu'une application directe du test unilatéral : on teste H_0 : *les deux distributions sont identiques*, contre H_1' : *les deux distributions diffèrent*, en déterminant R sur la suite obtenue en rangeant ensemble les observations des deux échantillons, puis en remplaçant chaque valeur par 0 ou 1 selon qu'elle provient du premier échantillon ou du second ; un nombre de runs « trop » faible peut indiquer une différence de positions (cas extrême: une suite comme 0000000011111111) ou de dispersions (exemple: 0000111111110000). Nous ne mentionnons ce test que pour être complets, car il semble moins puissant que les tests de SMIRNOV (cf. sect. 3.2) ou de CRAMÉR-von MISES (cf. sect. 3.3) sans pour autant être beaucoup plus simple à mettre en œuvre.

2.1.7 Exemple

Voici une suite de 13 observations dont on veut s'assurer qu'elle satisfait à l'hypothèse d'indépendance des observations (supposées équidistribuées) :

7 8 7 8 9 10 9 9 11 8 11 12 10

L'échantillon réordonné s'écrit : 7, 7, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 12, et la médiane vaut donc 9.

La suite se récrit ainsi (0000)(11)(0)(111), et donc $n_0 = n_1 = 5$ et $R = 4$.

Pour un test bilatéral de niveau $\alpha = 0,05$, on trouve dans la table $R^i = 2$ et $R^s = 10$. Comme $R = 4$, on ne peut pas rejeter l'hypothèse du caractère aléatoire de la suite observée.

2.1.8 Table de valeurs critiques du test du nombre de runs

On trouve dans chaque case de la table un ensemble de deux valeurs critiques (supérieure et inférieure), pour $n_0 < n_1$ (sinon, échanger les codes 0 et 1...) : dans la partie supérieure droite, on a les valeurs critiques du test bilatéral de niveau 5 % (ou, en ne conservant qu'une des deux valeurs, des tests unilatéraux de niveau 2,5 %) ; dans la partie inférieure gauche, celles

des tests unilatéraux de niveau 5 % (ou du test bilatéral de niveau 10 %). Un tiret indique qu'il n'y a pas de valeur critique.

Valeurs critiques du nombre de runs, niveau 5 % **(bilatéral)**

4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_0 / n_1
–	9	9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4
–	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	
	10	10	11	11	—	—	—	—	—	—	—	—	—	—	—	—	5
	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	
		11	12	12	13	13	13	13	—	—	—	—	—	—	—	—	6
		3	3	3	4	4	4	4	5	5	5	5	5	5	6	6	
4	8		13	13	14	14	14	14	15	15	15	—	—	—	—	—	7
	2		3	4	4	5	5	5	5	5	6	6	6	6	6	6	
5	9	9		14	14	15	15	16	16	16	16	17	17	17	17	17	8
	2	3		4	5	5	5	6	6	6	6	6	7	7	7	7	
6	9	10	11		15	16	16	16	17	17	18	18	18	18	18	18	9
	3	3	3		5	5	6	6	6	7	7	7	7	8	8	8	
7	9	10	11	12		16	17	17	18	18	18	19	19	19	20	20	10
	3	3	4	4		6	6	7	7	7	7	8	8	8	8	9	
8	—	11	12	13	13		17	18	19	19	19	20	20	20	21	21	11
	3	3	4	4	5		7	7	7	8	8	8	9	9	9	9	
9	—	11	12	13	14	14		19	19	20	20	21	21	21	22	22	12
	3	4	4	5	5	6		7	8	8	8	9	9	9	10	10	
10	—	11	12	13	14	15	16		20	20	21	21	22	22	23	23	13
	3	4	5	5	6	6	6		8	9	9	9	10	10	10	10	
11	—	—	13	14	15	15	16	17		21	22	22	23	23	23	24	14
	3	4	5	5	6	6	7	7		9	9	10	10	10	11	11	
12	—	—	13	14	15	16	17	17	18		22	23	23	24	24	25	15
	4	4	5	6	6	7	7	8	8		10	10	11	11	11	12	
13	—	—	13	14	15	16	17	18	18	19		23	24	25	25	25	16
	4	4	5	6	6	7	8	8	9	9		11	11	11	12	12	
14	—	—	13	14	16	17	17	18	19	20	20		25	25	26	26	17
	4	5	5	6	7	7	8	8	9	9	10		11	12	12	13	
15	—	—	—	15	16	17	18	19	19	20	21	21		26	26	27	18
	4	5	6	6	7	8	8	9	9	10	10	11		12	13	13	
16	—	—	—	15	16	17	18	19	20	21	21	22	23		27	27	19
	4	5	6	6	7	8	8	9	10	10	11	11	11		13	13	
17	—	—	—	15	16	17	18	19	20	21	22	22	23	24		28	20
	4	5	6	7	7	8	9	9	10	10	11	11	12	12		14	
18	—	—	—	15	16	18	19	20	21	21	22	23	24	24	25		
	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13		
19	—	—	—	15	16	18	19	20	21	22	23	23	24	25	25	26	
	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	—	—	—	15	17	18	19	20	21	22	23	24	25	25	26	27	27
	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15
n_1 / n_0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

(unilatéral)

2.2 Test de Cox et Stuart : tendance monotone

2.2.1 Définition du problème: hypothèses nulle et alternative

Déterminer si les observations de ce qui semble un échantillon ne présentent pas plutôt une position (moyenne ou médiane) à tendance monotone (croissant ou décroissant systématiquement).

H_0 : la position des observations successives est constante, contre l'une des alternatives :

H_1 : la position des observations successives présente une tendance monotone croissante ou décroissante (alternative bilatérale), ou bien :

H_1' : la position des observations successives présente une tendance croissante (alternative unilatérale), ou :

H_1'' : la position des observations successives présente une tendance décroissante (alternative unilatérale).

2.2.2 Conditions d'utilisation

Les données sont une suite $X_1, X_2, \ldots, X_n$ d'observations de variables aléatoires continues ou qualitatives ordonnées, indépendantes et de même loi sauf peut-être pour la position, qui peut être systématiquement croissante (ou décroissante).

2.2.3 Définition de la statistique, justification intuitive

Si la taille de l'échantillon n est paire, on utilise la suite de $(n/2)$ différences $(X_1 - X_{\frac{n}{2}+1}), (X_2 - X_{\frac{n}{2}+2}), \ldots, (X_{\frac{n}{2}} - X_n)$. Sous H_0, ces différences forment un échantillon d'une loi de médiane nulle, tandis que sous les alternatives, cette médiane est soit positive (tendance à la décroissance), soit négative (tendance croissante). Il suffit d'appliquer le test du signe à cet « échantillon » de taille $(n/2)$.

Si n est impair, on se ramène au cas pair en « éliminant » (en ignorant) l'observation du milieu, de numéro $(n+1)/2$.

2.2.4 Régions de rejet et de non-rejet

Cf. test du signe, sect. 1.3.4.

2.2.5 Mise en œuvre

Cf. test du signe, sect. 1.3.5.

2.2.6 Compléments

Les coefficients de corrélation de rangs de Spearman (cf. sect. 2.6) ou de Kendall (cf. sect. 2.7) calculés entre la série observée et la série des indices (dans le même ordre) peuvent constituer un moyen plus puissant (mais plus long à mettre en œuvre) de tester la présence d'une tendance monotone.

2.2.7 Exemple

(*Tiré de* SPRENT, 1992).

Le Ministère du Commerce américain publie des estimations, obtenues sur des échantillons indépendants, des distances parcourues chaque année aux États-Unis par des automobiles. Les chiffres sont les suivants, en milliers de miles, pour les années 1970 à 1983, dans cet ordre :

9,8 9,9 10,0 9,8 9,2 9,4 9,5 9,6 9,8 9,3 8,9 8,7 9,2 9,3

Prouvent-ils une tendance monotone pour l'une ou l'autre catégorie? Les 7 différences utilisées sont $(9,8-9,6)$, $(9,9-9,8)$, $(10,0-9,3)$, $(9,8-8,9)$, $(9,2-8,7)$, $(9,4-9,2)$ et $(9,5-9,3)$, et sont toutes positives. La région de rejet de niveau 0,05 (niveau réel 0,016) correspond à 0 ou 7 signes « plus ». La tendance monotone (bien sûr à la baisse) est donc prouvée.

2.2.8 Table

Cf. test du signe, sect. 1.3.8.

2.3 Test de Kolmogorov : ajustement d'une distribution théorique continue à une répartition empirique

2.3.1 Définition du problème: hypothèses nulle et alternative

Comparer globalement une distribution continue théorique $F_0(x)$ entièrement définie (de paramètres connus *a priori*, par exemple une loi uniforme $\mathcal{U}[0\ ;\ 10]$, ou une loi normale $\mathcal{N}(20,\ 2,7)$, etc.) et la distribution $F(x)$ inconnue, supposée continue, d'une population dont on observe un échantillon.

Avec le test bilatéral, on teste :

H_0 : pour tout x, $F(x) = F_0(x)$, contre H_1 : *il existe* x *tel que* $F(x) \neq F_0(x)$.

Avec les tests unilatéraux, on teste :

H_0' : *pour tout* x, $F(x) \geqslant F_0(x)$, contre H_1' : *il existe* x *tel que* $F(x) < F_0(x)$,

ou :

H_0'' : *pour tout* x, $F(x) \leqslant F_0(x)$, contre H_1'' : *il existe* x *tel que* $F(x) > F_0(x)$.

2.3.2 Conditions d'utilisation

Les données sont un échantillon $X_1, X_2, \ldots, X_n$ d'observations de variables aléatoires continues, indépendantes et de même loi.

On peut également utiliser le test pour des variables discrètes en catégories ordonnées, mais le test χ^2 d'ajustement est préférable avec des échantillons de grande taille.

La répartition théorique doit être entièrement définie, en particulier on ne doit pas estimer ses paramètres éventuels à partir de l'échantillon (car la distribution tabulée du test ne serait alors plus valide).

2.3.3 Définition de la statistique, justification intuitive

Soit $\widehat{F_n}(x) = (1/n).(\textit{nombre de valeurs inférieures ou égales à } x \textit{ dans l'échantillon})$ la répartition empirique de l'échantillon, dont on peut espérer qu'elle estime correctement la répartition inconnue $F(x)$ de la population étudiée, et soit $F_0(x)$ la répartition théorique testée.

La statistique de test est l'écart vertical maximal observé entre la répartition empirique $\widehat{F_n}$ et la répartition théorique F_0. Selon l'hypothèse testée, on utilisera $D = \sup_x |F_0(x) - \widehat{F_n}(x)|$ (test bilatéral), ou bien (tests unilatéraux) $D^+ = \sup_x [F_0(x) - \widehat{F_n}(x)]$ ou $D^- = \sup_x [\widehat{F_n}(x) - F_0(x)]$, qui devront prendre des valeurs plus grandes si l'hypothèse alternative correspondante est vérifiée (hypothèse nulle fausse).

2.3.4 Régions de rejet et de non-rejet

Elles sont définies par des valeurs critiques qui ne dépendent que de α et de n, mais pas de $F_0(x)$.

La statistique de test mesurant directement la différence (positive) entre la répartition observée et la répartition théorique, on rejettera H_0 lorsque cette différence excède la valeur critique $d_{1-\alpha,n}$ lue dans la table des quantiles (cf. sect. 2.3.8).

On rejette H_0 au profit de H_1 au niveau de signification α si $D > d_{1-\alpha,n}$, H_0' au profit de H_1' au niveau de signification α' si $D^+ > d_{1-\alpha',n}$, et H_0'' au profit de H_1'' au niveau de signification α' si $D^- > d_{1-\alpha',n}$.

2.3.5 Mise en œuvre

On commence par calculer la répartition théorique $F_0(x)$ et la répartition empirique $\widehat{F_n}(x)$ pour les valeurs de x observées dans l'échantillon : celle-ci est en effet une fonction en escalier, constante entre deux valeurs successives de ces observations.
On peut ensuite procéder :

- *graphiquement* : on porte sur un même graphe $F_0(x)$ et $\widehat{F_n}(x)$ et on lit directement D (ou D^+ ou D^-) qui est la plus grande distance verticale

entre les deux courbes (respectivement en allant de la première à la seconde respectivement du haut vers le bas, ou du bas vers le haut) ;

– *algébriquement* : on calcule les valeurs de $F_0(x)$ aux points observés x_i, puis toutes les différences $[F_0(x_i) - \widehat{F_n}(x_i)]$ et $[F_0(x_i) - \widehat{F_n}(x_{i-1})]$; D est alors la plus grande de ces différences en valeur absolue, D^+ est la plus grande valeur absolue des différences positives et D^- la plus grande des négatives.

2.3.6 Compléments

D'autres tests répondent au même type de problèmes :

– le test χ^2 d'ajustement (cf. sect. 2.4) est plus approprié pour des échantillons de grande taille (test asymptotique) si les variables sont discrètes ; il permet également des tests d'hypothèses *composées*, avec estimation de certains paramètres de la fonction de répartition, contrairement au test de KOLMOGOROV ;

– un test de CRAMÉR-VON MISES (non traité ici, ne pas confondre avec celui de la sect. 3.3), reposant sur les mêmes postulats que celui de KOLMOGOROV, prend en compte toutes les différences entre répartitions théorique et empirique, et non pas seulement la plus grande ; mais il ne traite que le cas bilatéral, et son éventuelle supériorité n'est d'ailleurs pas démontrée ;

– le test de LILLIEFORS (cf. sect. 2.5) permet, contrairement au test de KOLMOGOROV, de tester l'ajustement d'une loi normale dont les paramètres sont estimés sur les mêmes données.

2.3.7 Exemple

Considérons les longueurs de rupture d'un fil de 6 cm soumis à une contrainte. Nous voulons savoir si ces longueurs suivent une distribution uniforme $\mathcal{U}[0\ ;\ 6]$. Soit les observations suivantes de ce phénomène :

0,6	0,8	1,1	1,2	1,4	1,7	1,8	1,9	2,2	2,4
2,5	2,9	3,1	3,4	3,4	3,9	4,4	4,9	5,2	5,9

Nous calculons les valeurs des répartitions théorique et empirique correspondantes, ainsi que les écarts entre ces deux fonctions :

x_i	$F_0(x_i)$	$\widehat{F_n}(x_i)$	$F_0(x_i) - \widehat{F_n}(x_i)$	$F_0(x_i) - \widehat{F_n}(x_{i-1})$
0,6	0,10	0,05	0,05	*0,10*
0,8	0,13	0,10	0,03	0,08
1,1	0,18	0,15	0,03	0,08
1,2	0,20	0,20	0,00	0,05
1,4	0,23	0,25	−0,02	0,03
1,7	0,28	0,30	−0,02	0,03
1,8	0,30	0,35	−0,05	0,00
1,9	0,32	0,40	−0,08	−0,03
2,2	0,37	0,45	−0,08	−0,03
2,4	0,40	0,50	−0,10	−0,05
2,5	0,42	0,55	−0,13	−0,08
2,9	0,48	0,60	−0,12	−0,07
3,1	0,52	0,65	−0,13	−0,08
3,4	0,57	0,75	**−0,18**	−0,08
3,9	0,65	0,80	−0,15	*−0,10*
4,4	0,73	0,85	−0,12	−0,07
4,9	0,82	0,90	−0,08	−0,03
5,2	0,87	0,95	−0,08	−0,03
5,9	0,98	1,00	−0,02	0,03

Le plus grand écart est $D = 0,18 < d_{95\,\%,20} = 0,294$. On ne peut donc pas rejeter l'hypothèse H_0 que la longueur de rupture du fil suit une loi uniforme $\mathcal{U}[0\ ;\ 6]$.

2.3.8 Table des quantiles de la statistique de Kolmogorov

	Unilatéral :										
$\alpha' =$	0,900	0,950	0,975	0,990	0,995	$\alpha' =$	0,900	0,950	0,975	0,990	0,995
	Bilatéral :										
$\alpha =$	0,800	0,900	0,950	0,980	0,990	$\alpha =$	0,800	0,900	0,950	0,980	0,990
$n = 1$	0,900	0,950	0,975	0,990	0,995	$n = 21$	0,226	0,259	0,287	0,321	0,344
2	0,684	0,776	0,842	0,900	0,929	22	0,221	0,253	0,281	0,314	0,337
3	0,565	0,636	0,708	0,785	0,829	23	0,216	0,247	0,275	0,307	0,330
4	0,493	0,565	0,624	0,689	0,734	24	0,212	0,242	0,269	0,301	0,323
5	0,447	0,509	0,563	0,627	0,669	25	0,208	0,238	0,364	0,295	0,317
6	0,410	0,468	0,519	0,577	0,617	26	0,204	0,233	0,259	0,290	0,311
7	0,381	0,436	0,483	0,538	0,576	27	0,200	0,229	0,254	0,284	0,305
8	0,358	0,410	0,454	0,507	0,542	28	0,197	0,225	0,250	0,279	0,300
9	0,339	0,387	0,430	0,480	0,513	29	0,193	0,221	0,246	0,275	0,295
10	0,323	0,369	0,409	0,457	0,489	30	0,190	0,218	0,242	0,270	0,290
11	0,308	0,352	0,391	0,437	0,468	31	0,187	0,214	0,238	0,266	0,285
12	0,296	0,338	0,375	0,419	0,449	32	0,184	0,211	0,234	0,262	0,281

suite à la page suivante

... suite

	Unilatéral :										
$\alpha' =$	0,900	0,950	0,975	0,990	0,995	$\alpha' =$	0,900	0,950	0,975	0,990	0,995
	Bilatéral :										
$\alpha =$	0,800	0,900	0,950	0,980	0,990	$\alpha =$	0,800	0,900	0,950	0,980	0,990
13	0,285	0,325	0,361	0,404	0,432	33	0,182	0,208	0,231	0,258	0,277
14	0,275	0,314	0,349	0,390	0,418	34	0,179	0,205	0,227	0,254	0,273
15	0,266	0,304	0,338	0,377	0,404	35	0,177	0,202	0,224	0,251	0,269
16	0,258	0,295	0,327	0,366	0,392	36	0,174	0,199	0,221	0,247	0,265
17	0,250	0,286	0,318	0,355	0,381	37	0,172	0,196	0,218	0,244	0,262
18	0,244	0,279	0,309	0,346	0,371	38	0,170	0,194	0,215	0,241	0,258
19	0,237	0,271	0,301	0,337	0,361	39	0,168	0,191	0,213	0,238	0,255
20	0,232	0,265	0,294	0,329	0,352	40	0,165	0,189	0,210	0,235	0,252

Pour $n > 40$, on utilisera, pour les mêmes niveaux de signification, les approximations suivantes (respectivement) : $1{,}07/\sqrt{n}$, $1{,}22/\sqrt{n}$, $1{,}36/\sqrt{n}$, $1{,}52/\sqrt{n}$, $1{,}63/\sqrt{n}$.

2.4 Test χ^2 d'ajustement d'une distribution théorique à des fréquences observées

2.4.1 Définition du problème: hypothèses nulle et alternative

Comparer globalement une distribution théorique (dont *a priori* certains paramètres peuvent être inconnus) et la distribution (inconnue) d'une population dont on observe un échantillon.

On considère un échantillon aléatoire et simple d'une population infinie dont les individus sont répartis en catégories ou classes. Le but est de tester si la population dont est issu l'échantillon possède une fonction de répartition donnée (par exemple lois uniformes, binomiales, de POISSON, etc.).

Soit F la fonction de répartition inconnue de la population dont est issu l'échantillon, et F_0 la fonction de répartition théorique à laquelle on veut la comparer. Nous devons considérer deux cas selon que celle-ci est complètement définie ou non :

- hypothèse simple (loi théorique entièrement définie) :

$$H_0 \ : \ F = F_0 \text{ contre } H_1 \ : \ F \neq F_0$$

- hypothèse composée (k paramètres θ de la loi théorique sont inconnus) :

$H_0 \ : \ F \in \{F(\theta) \ ; \theta \text{ quelconque}\}$ contre $H_1 \ : \ F \notin \{F(\theta) \ ; \theta \text{ quelconque}\}$

L'hypothèse simple est associée à une distribution parfaitement définie, par exemple $X_i \sim \mathcal{N}(2,5)$, alors que l'hypothèse composée est associée à une fonction de répartition dont seule la forme est connue (par exemple, $X_i \sim \mathcal{N}(\mu, \sigma^2)$ avec μ et σ^2 inconnus).

Il est clair qu'ici la conclusion espérée est l'acceptation de H_0, et que la puissance du test joue donc un rôle important.

2.4.2 Conditions d'utilisation

Les données sont constituées par un n-échantillon $X_1, X_2, \ldots, X_n$ de variables aléatoires indépendantes et équidistribuées. Les observations de ces variables sont réparties, en fonction d'un critère qualitatif ou quantitatif, en c classes ou cellules $C_1, C_2, \ldots, C_c$ et les nombres d'observations dans chacune d'elles sont rangés sous forme de tableau de contingence $1 \times c$:

Classes :	C_1	C_2	$\ldots$	C_c	Total
Effectifs observés :	o_1	o_2	$\ldots$	o_c	n

Ce test peut être utilisé pour des données à valeurs continues regroupées arbitrairement en classes, surtout pour les tests d'hypothèses composées, mais dans ce cas il y a perte d'information (cf. sect. 2.4.6 Compléments).

Soit p_j la probabilité, sous H_0, qu'une variable X_i quelconque appartienne à la classe C_j. Lorsque X_i est continue,

$$p_j = \int_{C_j} f_\theta(x)\, dx,$$

et lorsque X_i est discrète,

$$p_j = \sum_{x \in C_j} \Pr_\theta\{X_j = x\}.$$

Le test n'est valide qu'asymptotiquement, en pratique on accepte l'approximation, lorsque $c = 2$, si np_1 et np_2 sont supérieurs à 5, et lorsque $c > 2$, si plus de 75 % des np_j sont supérieurs à 5 et si aucun des np_j n'est inférieur à 1. Si ces conditions ne sont pas remplies, il faut soit regrouper des classes (si cela a un sens), soit réaliser plus d'observations, soit renoncer au test.

2.4.3 Définition de la statistique, justification intuitive

Pour les tests d'hypothèses simples, la statistique de test est :

$$\chi^2 = \sum_{j=1}^{j=c} \frac{(o_j - np_j)^2}{np_j}$$

C'est donc une façon de mesurer l'écart global entre les effectifs observés o_j et les effectifs np_j prédits par la distribution théorique, en pondérant chaque carré d'écart d'une classe par son effectif théorique (son importance).

Pour les tests d'hypothèses composées, où k paramètres θ de la distribution théorique sont inconnus, on procède de même après avoir remplacé, dans le calcul des p_j qui sont des fonctions des θ, les θ inconnus par leurs estimations par maximum de vraisemblance $\widehat{\theta}$.

2.4.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $\chi^2 \geqslant \chi^2_{\alpha,c-1-k}$, le $(1-\alpha)$-quantile de la distribution χ^2 à $c-1-k$ degrés de liberté que l'on trouve dans la table (cf. sect. 2.4.8, p. 34) ; dans le cas du test d'une hypothèse simple, on pose k (nombre de paramètres inconnus et estimés) égal à 0.

Remarques : ce test est toujours « bilatéral ». Il est conservatif (trop sévère) avec des hypothèses composées. L'alternative étant très vaste, il est difficile d'évaluer la puissance du test.

2.4.5 Mise en œuvre

Calcul de la statistique
On calcule d'abord les probabilités théoriques p_j (éventuellement après calcul des estimations $\widehat{\theta}$ des paramètres inconnus), puis les effectifs théoriques np_j, et la statistique se calcule facilement avec la formule de la section 2.4.3, ou avec la formule équivalente suivante : $\chi^2 = (\sum_{j=1}^{j=c} o_j^2/np_j) - n$.

Distribution sous H_0
La variable χ^2 sous l'hypothèse nulle suit asymptotiquement une distribution χ^2, à $c-1$ degrés de liberté en cas d'hypothèse simple ($c-1$ car les $c-1$ premières fréquences o_j et np_j suffisent à déterminer complètement la dernière, par exemple $o_c = n - \sum_{j=1}^{j=c-1} o_j$), ou à $c-1-k$ degrés de liberté en cas d'hypothèse composée avec k paramètres estimés sur l'échantillon.

2.4.6 Compléments

Pour l'ajustement de distributions continues complètement spécifiées, et surtout lorsque l'on dispose de peu d'observations, le test de Kolmogorov est mieux adapté.

Il existe des tests spécifiques pour certaines familles de distributions, comme pour les lois normales (avec estimation des paramètres), le test de Lilliefors (cf. sect. 2.5), ou pour les lois de Poisson des tests faisant intervenir le rapport variance/moyenne (cf. Cochran, 1954 : *Biometrics* **10**, 417-451, ou Rao, 1956 : *Biometrics* **12**, 264-282).

2.4.7 Exemples

Test d'hypothèse simple : lois de Mendel
Le croisement de deux types de maïs a donné en deuxième génération 1 301 plantes de quatre phénotypes : 773 vertes, 231 or, 238 vertes rayées et 59 vert et or rayées. On veut tester si les données correspondent aux prédictions du

schéma de MENDEL : $p_1 = 9/16$, $p_2 = 3/16 = p_3$ et $p_4 = 1/16$.

j	o_j	np_j	$o_j - np_j$	$(o_j - np_j)^2$	$(o_j - np_j)^2/np_j$
1	773	731,9	41,1	1 689,21	2,31
2	231	243,9	−12,9	166,41	0,68
3	238	243,9	−5,9	34,81	0,14
4	59	81,3	−22,3	497,29	6,12
$\sum$	1 301	1 301	0		$\chi^2 = 9,25$

La valeur seuil $\chi^2_{0,95,\,4-1}$ vaut 7,82. La valeur de la statistique pour l'échantillon étant plus élevée, nous rejetons ici l'hypothèse H_0 : le schéma de MENDEL ne s'applique pas aux ségrégations observées.

Test d'hypothèse composée : ajustement d'une loi de Poisson
Dans un hématimètre à 400 carrés, 1872 organismes ont été observés, répartis de la façon suivante :

Nombre par case :	0 ou 1	2	3	4	5	6	7	8	9	10 ou plus	Total
Nombre de cases :	20	43	53	86	70	54	37	18	10	9	400

Nous voulons savoir si les organismes sont répartis au hasard. Si c'est le cas, le nombre observé par carré suit une loi multinomiale, approchée par une loi de POISSON. Nous testons donc $H_0 \;:\; F \in \{\mathcal{P}(\lambda)\ ;\lambda \text{ réel positif}\}$ contre l'alternative contraire.

L'estimation du maximum de vraisemblance du paramètre d'une loi de POISSON est donnée par la moyenne de l'échantillon, $\widehat{\lambda} = 1\,872/400 = 4,68$.

On calcule les p_j avec la formule de la loi de POISSON : $p_j = e^{-\widehat{\lambda}}(\widehat{\lambda}^j/j\,!)$.

j	o_j	$p_j(\widehat{\lambda})$	$np_j(\widehat{\lambda})$	o_j^2	$o_j^2/np_j(\widehat{\lambda})$
0 ou 1	20	0,053	21,2	400	18,87
2	43	0,102	40,8	1849	45,31
3	53	0,159	63,6	2809	44,95
4	86	0,185	74,0	7396	99,95
5	70	0,174	69,6	4900	70,40
6	54	0,135	54,0	2916	54,00
7	37	0,090	36,0	1369	38,03
8	18	0,053	21,2	324	15,28
9	10	0,028	11,2	100	8,93
10 ou plus	9	0,021	8,4	81	9,64
$\sum$	$n = 400$	1	$n = 400$		404,58

On obtient $\chi^2 = 4,58$, que l'on compare au $\chi^2_{0,95,\,10-1-1} = 15,51$: nous ne pouvons donc pas rejeter l'hypothèse nulle.

2.4.8 Table de quantiles $\chi^2_{1-\alpha,\,\nu}$ des distributions χ^2

$1-\alpha$	0,750	0,900	0,950	0,975	0,990	0,995	0,999
$\nu = 1$	1,323	2,706	3,841	5,024	6,635	7,879	10,83
2	2,773	4,605	5,991	7,378	9,210	10,60	13,82
3	4,108	6,251	7,815	9,348	11,34	12,84	16,27
4	5,385	7,779	9,488	11,14	13,28	14,86	18,47
5	6,626	9,236	11,07	12,83	15,09	16,75	20,51
6	7,841	10,4	12,59	14,45	16,81	18,55	22,46
7	9,037	12,02	14,07	16,01	18,48	20,28	24,32
8	10,22	13,36	15,51	17,53	20,09	21,96	26,13
9	11,39	14,68	16,92	19,02	21,67	23,59	27,88
10	12,55	15,99	18,31	20,48	23,21	25,19	29,59
11	13,70	17,28	19,68	21,92	24,73	26,76	31,26
12	14,85	18,55	21,03	23,34	26,22	28,30	32,91
13	15,98	19,81	22,36	24,74	27,69	29,82	34,53
14	17,12	21,06	23,68	26,12	29,14	31,32	36,12
15	18,25	22,31	25,00	27,49	30,58	32,80	37,70
16	19,37	23,54	26,30	28,85	32,00	34,27	39,25
17	20,49	24,77	27,59	30,19	33,41	35,72	40,79
18	21,60	25,99	28,87	31,53	34,81	37,16	42,31
19	22,72	27,20	30,14	32,85	36,19	38,58	43,82
20	23,83	28,41	31,41	34,17	37,57	40,00	45,32
21	24,93	29,62	32,67	35,48	38,93	41,40	46,80
22	26,04	30,81	33,92	36,78	40,29	42,80	48,27
23	27,14	32,01	35,17	38,08	41,64	44,18	49,73
24	28,24	33,20	36,42	39,37	42,98	45,56	51,18
25	29,34	34,38	37,65	40,65	44,31	46,93	52,62
26	30,43	35,56	38,89	41,92	45,64	48,29	54,05
27	31,53	36,74	40,11	43,19	46,96	49,64	55,48
28	32,62	37,92	41,34	44,46	48,28	50,99	56,89
29	33,71	39,09	42,56	45,72	49,59	52,34	58,30
30	34,80	40,26	43,77	46,98	50,89	53,67	59,70
40	45,62	51,81	55,76	59,34	63,69	66,77	73,40
50	56,33	63,17	67,50	71,42	76,15	79,49	86,66
60	66,98	74,40	79,08	83,30	88,38	91,95	99,61
70	77,58	85,53	90,53	95,02	100,4	104,2	112,3
80	88,13	96,58	101,9	106,6	112,3	116,3	124,8
90	98,65	107,6	113,1	118,1	124,1	128,3	137,2
100	109,1	118,5	124,3	129,6	135,8	140,2	149,4
$z_{1-\alpha}$	0,675	1,282	1,645	1,960	2,326	2,576	3,090

Lorsque $\nu > 100$, on peut utiliser une approximation de $\chi^2_{1-\alpha,\,\nu}$ en fonction
des quantiles $z_{1-\alpha}$ de la loi de GAUSS : $\chi^2_{1-\alpha,\,\nu} \approx (z_{1-\alpha} + \sqrt{2\nu - 1})^2/2$.

2.5 Test de Lilliefors : ajustement d'une loi de Gauss

2.5.1 Définition du problème: hypothèses nulle et alternative

Vérifier si la distribution (inconnue) d'une population n'est pas trop éloignée de la famille des lois de GAUSS. On teste :

H_0 : $F \in \{\mathcal{N}(\mu, \sigma^2)$; μ et σ quelconques$\}$, contre
H_1 : $F \notin \{\mathcal{N}(\mu, \sigma^2)$; μ et σ quelconques$\}$.

2.5.2 Conditions d'utilisation

Les données sont un n-échantillon $X_1, X_2, \ldots, X_n$ de variables aléatoires continues, indépendantes et de même répartition inconnue $F(x)$.

2.5.3 Définition de la statistique, justification intuitive

On utilise la moyenne empirique $\overline{X} = (\sum_{i=1}^{i=n} X_i)/n$ et l'écart-type empirique $\widehat{\sigma}$ défini par $\widehat{\sigma}^2 = \sum_{i=1}^{i=n}(X_i - \overline{X})^2/(n-1)$ pour « standardiser » les données en définissant les variables $Z_i = (X_i - \overline{X})/\widehat{\sigma}$.

Soit $\widehat{F}_n(x)$ la fonction de répartition empirique de ces Z_i, c.à-d. ù

$$\widehat{F}_n(x) = \frac{1}{n} \times \text{ nombre de variables } Z_i \leq x.$$

Et soit $\Phi(x)$ la fonction de répartition de la loi de GAUSS centrée et réduite $\mathcal{N}(0,\ 1)$. La statistique de test est analogue à celle du test bilatéral de KOLMOGOROV : c'est le plus grand écart vertical entre les deux, $D = \sup_x |\Phi(x) - \widehat{F}_n(x)|$.

Remarquer que si la définition de la statistique est identique à celle de KOLMOGOROV, sa distribution ne peut être identique car les estimateurs de la moyenne et de l'écart-type introduisent, par rapport à la situation du test de KOLMOGOROV, un aléa supplémentaire dont il faut tenir compte.

2.5.4 Régions de rejet et de non-rejet

On rejette l'hypothèse H_0 au profit de H_1 pour un niveau de signification α lorsque $D > d_{\alpha,\,n}$, le $(1-\alpha)$-quantile de la statistique de LILLIEFORS lu dans la table (cf. sect. 2.5.8).

2.5.5 Mise en œuvre

Sauf pour la table, différente, des valeurs critiques, la mise en œuvre est identique à celle du test de KOLMOGOROV une fois calculées les Z_i.

2.5.6 Compléments

Il existe de nombreux tests de normalité et l'alternative très vaste rend délicates les comparaisons. Citons le test de SHAPIRO et WILK (cf. p. ex. CONOVER, *Practical Nonparametric Statistics*, Wiley & Sons, New-York, 2^e éd., 1980, sect. 6.2), et des tests fondés sur les coefficients de symétrie et d'aplatissement (cf. BOWMAN et SHENTON, 1975, *Biometrika* **62**, 243-250).

2.5.7 Exemple

(*Tiré de* SPRENT, 1992).

Dans le cimetière de Badenscallie, en Écosse, on a relevé l'âge de décès des hommes de quatre clans. Parmi les 117 âges notés, on a tiré un échantillon aléatoire de 30 valeurs :

11	13	14	22	29	30	41	41	52	55	56	59	65	65	66
74	74	75	77	81	82	82	82	82	83	85	85	87	87	88

Est-il raisonnable de penser que les âges de décès suivent une loi normale?

x_i	Z_i	$\Phi(Z_i)$	$\widehat{F}_n(Z_i)$	$\Phi(Z_i) - \widehat{F}_n(Z_i)$	$\Phi(Z_i) - \widehat{F}_n(Z_{i-1})$
11	−2,014	0,022	0,033	−0,011	0,022
13	−1,934	0,026	0,067	−0,044	−0,007
14	−1,894	0,029	0,100	−0,071	−0,038
22	−1,575	0,058	0,133	−0,075	−0,042
29	−1,295	0,098	0,167	−0,069	−0,035
30	−1,255	0,105	0,200	−0,095	−0,062
41*	−0,816	0,207	0,267	−0,060	−0,007
52	−0,377	0,353	0,300	0,053	0,086
55	−0,257	0,399	0,333	0,066	0,099
56	−0,217	0,414	0,367	0,047	0,081
59	−0,097	0,461	0,400	0,061	0,094
65*	0,142	0,556	0,467	0,089	0,156
66	0,183	0,572	0,500	0,072	0,105
74*	0,502	0,692	0,567	0,125	**0,192**
75	0,542	0,706	0,600	0,106	0,139
77	0,622	0,733	0,633	0,100	0,133
81	0,781	0,782	0,667	0,115	0,149
82*	0,821	0,794	0,800	−0,006	0,127
83	0,861	0,805	0,833	−0,028	−0,005
85*	0,942	0,827	0,900	−0,073	−0,006
87*	1,021	0,846	0,967	−0,121	−0,054
88	1,061	0,856	1,000	*−0,144*	−0,111

(l'astérisque* signale les ex æquo, expliquant des sauts plus importants de $\widehat{F_n}$).

La moyenne et l'écart-type de cet échantillon valent respectivement 61,43 et 25,04, ce qui permet de calculer les Z_i ainsi que les écarts entre leur répartition empirique et la répartition gaussienne standard.

La plus grande différence observée entre les deux fonctions de répartition vaut 0,192, qui est supérieur à la valeur critique $d_{0,05,30} = 0,161$ de la statistique de LILLIEFORS : nous rejetons donc l'hypothèse nulle de normalité de ces âges de décès.

2.5.8 Table de quantiles de la statistique de Lilliefors

$1-\alpha$	0,80	0,85	0,90	0,95	0,99
$n=4$	0,300	0,319	0,352	0,381	0,417
5	0,285	0,299	0,315	0,337	0,405
6	0,265	0,277	0,294	0,319	0,364
7	0,247	0,258	0,276	0,300	0,348
8	0,233	0,244	0,261	0,285	0,331
9	0,223	0,233	0,249	0,271	0,311
10	0,215	0,224	0,239	0,258	0,294
11	0,206	0,217	0,230	0,249	0,284
12	0,199	0,212	0,223	0,242	0,275
13	0,190	0,202	0,214	0,234	0,268
14	0,183	0,194	0,207	0,227	0,261
15	0,177	0,187	0,201	0,220	0,257
16	0,173	0,182	0,195	0,213	0,250
17	0,169	0,177	0,189	0,206	0,245
18	0,166	0,173	0,184	0,200	0,239
19	0,163	0,169	0,179	0,195	0,235
20	0,160	0,166	0,174	0,190	0,231
25	0,142	0,147	0,158	0,173	0,200
30	0,131	0,136	0,144	0,161	0,187

Pour $n > 30$, on utilisera les approximations suivantes, pour les mêmes niveaux de signification (respectivement) : $0,736/\sqrt{n}$, $0,768/\sqrt{n}$, $0,805/\sqrt{n}$, $0,886/\sqrt{n}$, et $1,031/\sqrt{n}$.

2.6 Test de corrélation de Spearman

2.6.1 Définition du problème: hypothèses nulle et alternative

On considère un échantillon de n observations de deux variables X et Y. Tester l'indépendance de X et Y :

$$H_0 = \{X \text{ et } Y \text{ sont indépendants}\}.$$

Hypothèse bilatérale :

$$H_1 = \{X \text{ et } Y \text{ sont corrélées}\}.$$

Hypothèses unilatérales :

$$H_1' = \{X \text{ et } Y \text{ sont corrélées positivement}\}.$$

$$H_1'' = \{X \text{ et } Y \text{ sont corrélées négativement}\}.$$

2.6.2 Conditions d'utilisation

Ce test ne peut pas s'utiliser avec de très petits échantillons ($n < 4$). Il faut avoir des variables quantitatives ou qualitatives ordinales appariées $(x_1, y_1), \ldots, (x_n, y_n)$.

2.6.3 Définition de la statistique, justification intuitive

On calcule la statistique de test ρ basée sur les rangs des variables. On note (r_i, s_i) les rangs de (x_i, y_i). S'il y a des ex-æquo, on prendra les rangs moyens.

$$\rho = \frac{\sum_{i=1}^{n} r_i s_i - C}{\sqrt{\left[\sum_{i=1}^{n} r_i^2 - C\right] \times \left[\sum_{i=1}^{n} s_i^2 - C\right]}}$$

avec $C = n(n+1)^2/4$.

S'il n'y a pas d' ex æquo :

$$\rho = 1 - \frac{6T}{n(n^2-1)} \quad \text{avec } T = \sum_{i=1}^{n} (r_i - s_i)^2.$$

On a

$$-1 \leq \rho \leq +1$$

Si $\rho \simeq 1$ alors X et Y sont corrélées positivement, si $\rho \simeq -1$ alors X et Y sont corrélées négativement et si $\rho \simeq 0$ alors X et Y sont indépendantes.

2.6.4 Régions de rejet et de non-rejet

On rejette H_0 si $|\rho| \geq \rho_\alpha$.

2.6.5 Mise en œuvre

Calcul de la statistique Il ne présente pas de difficultés.

Distribution sous H_0 La distribution de ρ sous H_0 est tabulée (cf Table A.9 Sprent).

Pour $n > 20$, ρ peut être assimilé à une loi normale $\mathcal{N}(0, \frac{1}{n-1})$, c'est-à-dire $\rho = \frac{z}{\sqrt{n-1}}$ où z est suit la loi $\mathcal{N}(0,1)$. Cette approximation est satisfaisante lorsque $n > 30$.

2.6.6 Compléments

En l'absence d'ex-æquo, on peut utiliser la statistique T et le test de Hotelling-Pabst. (cf Conover : table 9)

2.6.7 Mise en œuvre informatique

L'exemple traité ci-dessous est tiré de Sprent (1992).

Dans une expérience de pharmacologie sur des agents β-bloquants, Sweeting (1982) a enregistré, sur un groupe de chiens, la consommation cardiaque d'oxygène (CCO) et la pression ventriculaire gauche (PVG).

Chien	A	B	C	D	E	F	G
CCO	78	92	116	90	106	78	99
PVG	32	33	45	30	38	24	44

SPLUS On peut utiliser la fonction `cor.test`.

```
CCO_c(78,92,116,90,106,78,99)
PVG_c(32,33,45,30,38,24,44)

cor.test(CCO,PVG,alternative="two.sided",method="s")
```

On obtient alors comme sortie

```
      Spearman's rank correlation

data:  CCO and PVG
normal-z = 2.1652, p-value = 0.0304
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
 0.9017857
```

Splus utilise la statistique suivante :

$$z = \rho\sqrt{n-1} - \frac{6\sqrt{n-1}}{n^3 - n}.$$

SAS Les commandes à utiliser sont

```
data chiens;
input numchien $ CCO PVG;
cards;
A 78  32
B 92  33
C 116 45
D 90  30
E 106 38
F 78  24
G 99  44
;

proc corr data=chiens spearman;
  var CCO PVG;
  run;
quit;
```

On obtient alors comme sortie

```
                        Correlation Analysis

                    2 'VAR' Variables:  CCO      PVG


                         Simple Statistics

Variable    N     Mean  Std Dev   Median Minimum  Maximum

CCO         7    94.14    14.05    92.00   78.00   116.00
PVG         7    35.14     7.63    33.00   24.00    45.00


Spearman Correlation Coefficients / Prob > |R| under Ho: Rho=0
/ N = 7

                                    CCO              PVG

                CCO            1.00000          0.90094
                                0.0              0.0056

                PVG            0.90094          1.00000
                                0.0056           0.0
```

2.7 Test de corrélation de Kendall

2.7.1 Définition du problème: hypothèses nulle et alternative

On considère un échantillon de n observations de deux variables X et Y. On note c la corrélation entre X et Y. Tester l'indépendance de X et Y :

$$H_0 = \{X \text{ et } Y \text{ sont indépendants}\}.$$

Hypothèse bilatérale :

$$H_1 = \{X \text{ et } Y \text{ sont corrélés}\}.$$

Hypothèses unilatérales :

$$H_1' = \{X \text{ et } Y \text{ sont corrélés positivement}\}.$$

$$H_1'' = \{X \text{ et } Y \text{ sont corrélés négativement}\}.$$

2.7.2 Conditions d'utilisation

Ce test ne peut pas s'utiliser avec de très petits échantillons ($n < 4$). Il faut avoir des variables quantitatives ou qualitatives ordinales appariées $(x_1, y_1), \ldots, (x_n, y_n)$.

2.7.3 Définition de la statistique, justification intuitive

On calcule la statistique de test τ basée sur la concordance des variables.

Une paire d'observations (x_i, y_i) et (x_j, y_j) est concordante si le rapport $(y_j - y_i)/(x_j - x_i)$ est positif et discordante sinon.

$$\tau = \frac{n_\mathrm{c} - n_\mathrm{d}}{n(n-1)/2}$$

où $n_\mathrm{c}, n_\mathrm{d}$ sont respectivement le nombre de paires concordantes et discordantes. On a

$$-1 \leq \tau \leq +1.$$

Si $\tau \simeq 1$ alors X et Y sont corrélées positivement, si $\tau \simeq -1$ alors X et Y sont corrélées négativement et si $\tau \simeq 0$ alors X et Y sont indépendantes.

2.7.4 Régions de rejet et de non-rejet

On note $S = n_c - n_d$ la statistique de test.

On rejette H_0 si $|S| \geq S_\alpha$.

2.7.5 Mise en œuvre

Calcul de la statistique La méthode calculatoire ne présente pas de difficultés.

Présentation graphique du calcul de S On note sur deux rangs parallèles les valeurs de x_i rangées par ordre croissant sur la première ligne et on associe sur la deuxième ligne les valeurs de y_i correspondantes.

Ensuite, on relie les X et Y de rangs homologues.

Le nombre n_d est le nombre d'intersections obtenues en évitant les intersections superposées.

S'il n'y a pas d'ex æquo, on sait que $n_c + n_d = n(n-1)/2$ et on détermine ainsi n_c. Si on note t le nombre de paires où on a des ex æquo, on utilise la relation :

$$n_c + n_d + t = n(n-1)/2.$$

Application (cf. exemple section 2.6.7)

78	78	90	92	99	106	116
24	32	30	33	44	38	45

En reliant les variables X et Y de rangs homologues, on obtient deux intersections et ainsi $n_d = 2$. D'autre part, on a $t = 2$ car on a 2 paires avec des ex-æquo (la paire $((78,24),(78,32))$ et son inverse $((78,32),(78,24))$). En utilisant la relation $n_c + n_d + t = n(n-1)/2$, on a :

$$n_c = 18 \text{ et } \tau = 15/21.$$

Distribution sous H_0 La statistique S est tabulée sous H_0 pour $n \leq 20$. (cf Table A.10 Sprent)

Pour $n \geq 10$, la loi de distribution de τ sous H_0 est approximée par la loi normale de moyenne $\mu = 0$ et d'écart type

$$\sigma = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

Remarque : Splus fournit la statistique $z = \frac{\tau}{\sigma}$.

2.7.6 Table

Table A.10 Sprent.

2.7.7 Mise en œuvre informatique

Les données traitées sont les mêmes que dans la section précédente.

SPLUS On utilise la fonction `cor.test`.

```
CCO_c(78,92,116,90,106,78,99)
PVG_c(32,33,45,30,38,24,44)

cor.test(CCO,PVG,alternative="two.sided",method="k")
```

On obtient alors en sortie

```
        Kendall's rank correlation tau

data:  CCO and PVG
normal-z = 2.2528, p-value = 0.0243
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
 0.7142857
```

Splus utilise la statistique suivante

$$z = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}.$$

SAS On utilise les commandes

```
data chiens;
input numchien $ CCO PVG;
cards;
A 78  32
B 92  33
C 116 45
D 90  30
E 106 38
F 78  24
G 99  44
;

proc corr data=chiens kendall;
  var CCO PVG;
  run;
quit;
```

pour obtenir

```
                        Correlation Analysis

                    2 'VAR' Variables:  CCO      PVG


                         Simple Statistics

Variable     N     Mean Std Dev   Median Minimum  Maximum
```

```
CCO          7    94.14   14.05    92.00   78.00   116.00
PVG          7    35.14    7.63    33.00   24.00    45.00


Kendall Correlation Coefficients / Prob > |R| under Ho: Rho=0
/ N = 7

                                CCO             PVG

                CCO          1.00000         0.78072
                              0.0             0.0151

                PVG          0.78072         1.00000
                             0.0151           0.0
```

2.8 Test de corrélation de Pearson

2.8.1 Définition du problème: hypothèses nulle et alternative

On considère un échantillon de n observations de deux variables X et Y. Tester l'indépendance de X et Y :

$$H_0 = \{X \text{ et } Y \text{sont indépendants}\}.$$

Hypothèse bilatérale :

$$H_1 = \{X \text{ et } Y \text{sont corrélés}\}.$$

Hypothèses unilatérales :

$$H_1' = \{X \text{ et } Y \text{sont corrélés positivement}\}.$$

ou

$$H_1'' = \{X \text{ et } Y \text{sont corrélés négativement}\}.$$

2.8.2 Conditions d'utilisation

Ce test ne peut pas s'utiliser avec de très petits échantillons ($n < 4$). Il faut avoir des variables quantitatives ou qualitatives ordinales appariées $(x_1, y_1), \ldots, (x_n, y_n)$.

2.8.3 Définition de la statistique, justification intuitive

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

On a

$$-1 \leq r \leq +1.$$

Si $r \simeq 1$ alors les points de coordonnées (x_i, y_i) sont presque alignés sur une droite de pente positive, si $r \simeq -1$ alors ils sont presque alignés sur une droite de pente négative et si $r \simeq 0$ alors les points ne présentent pas de liaison linéaire.

Sans le postulat de normalité bivariate, les tests d'hypothèses et d'estimation avec r dépendent de la distribution du couple (x,y).

2.8.4 Mise en œuvre

Calcul de la statistique Ne présente pas de difficultés.

Distribution sous H_0 Avec le postulat de normalité bivariate, sous H_0, on utilise la statistique $t = r\sqrt{\frac{n-2}{1-r^2}}$ où t suit une loi de Student à $n-2$ degrés de liberté.

2.8.5 Mise en œuvre informatique

Les données traitées sont les mêmes que dans la section précédente.

SPLUS La fonction à utiliser est `cor.test`.

```
CCO_c(78,92,116,90,106,78,99)
PVG_c(32,33,45,30,38,24,44)

cor.test(CCO,PVG,alternative="two.sided",method="p")
```

On obtient comme sortie

```
        Pearson's product-moment correlation

data:  CCO and PVG
t = 3.6655, df = 5, p-value = 0.0145
alternative hypothesis: true coef is not equal to 0
sample estimates:
       cor
 0.8536946
```

Splus utilise la statistique

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

SAS Les commandes à utiliser sont

```
data chiens;
input numchien $ CCO PVG;
cards;
A 78  32
B 92  33
C 116 45
D 90  30
E 106 38
F 78  24
G 99  44
;

proc corr data=chiens pearson;
  var CCO PVG;
  run;
quit;
```

On obtient comme sortie

```
                          Correlation Analysis

                    2 'VAR' Variables:  CCO      PVG


                          Simple Statistics

Variable     N     Mean Std Dev   Median Minimum  Maximum

CCO          7    94.14   14.05    92.00   78.00   116.00
PVG          7    35.14    7.63    33.00   24.00    45.00


Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0
/ N = 7

                                     CCO              PVG

                  CCO            1.00000          0.85369
                                  0.0              0.0145

                  PVG            0.85369          1.00000
                                  0.0145           0.0
```

Deuxième partie

Deux échantillons indépendants

Chapitre 1

Comparaison des positions

1.1 Test t de Student

1.1.1 Définition du problème: hypothèses nulle et alternative

Tester l'égalité des moyennes des deux populations d'où sont tirés les deux échantillons X_i, $i = 1, \ldots, m$, de moyenne μ_X, et Y_j, $j = 1, \ldots, n$, de moyenne μ_Y, tous deux de variance σ^2, avec les trois paramètres inconnus :

H_0 : $\mu_X = \mu_Y$ contre H_1 : $\mu_X \neq \mu_Y$ (alternative bilatérale), ou bien :

H'_0 : $\mu_X \leqslant \mu_Y$ contre H'_1 : $\mu_X > \mu_Y$ (alternative unilatérale), ou :

H''_0 : $\mu_X \geqslant \mu_Y$ contre H''_1 : $\mu_X < \mu_Y$ (alternative unilatérale).

1.1.2 Conditions d'utilisation

Il faut avoir un m-échantillon de variables continues $X_1, \ldots, X_m$ et un n-échantillon de variables continues $Y_1, \ldots, Y_n$. Les variables X_i et Y_j doivent impérativement être indépendantes et de même variance.

De plus, le test n'est exact en toute rigueur que si les distributions des variables X et Y sont gaussiennes : $X_i \sim \mathcal{N}(\mu_X,\ \sigma^2)$ et $Y_j \sim \mathcal{N}(\mu_Y,\ \sigma^2)$. Il reste cependant approximativement valide pour d'autres distributions, à condition d'avoir des tailles d'échantillons suffisantes (à partir de 30 environ).

1.1.3 Définition de la statistique, justification intuitive

On définit t à partir des statistiques $\overline{X} = (\sum_{i=1}^{m} X_i)/m$ estimant μ_X, $\overline{Y} = (\sum_{j=1}^{n} Y_j)/n$ estimant μ_Y, et, pour l'écart-type σ, la statistique $\widehat{\sigma}$ définie par $\widehat{\sigma}^2 = [\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2]/(m + n - 2)$, par la formule :

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\widehat{\sigma}\sqrt{(1/m) + (1/n)}}$$

où $(\mu_X - \mu_Y)$ ne figure que pour mémoire (cette différence est *nulle* sous H_0) et pour montrer que l'on mesure ainsi l'écart entre la différence de moyennes observée $(\overline{X} - \overline{Y})$ et celle que l'on teste $(\mu_X - \mu_Y)$, en tenant compte, avec l'écart-type mesuré $\widehat{\sigma}$, de la variabilité propre du phénomène.

1.1.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $|t| > c_{\alpha/2}$, ou bien :
On rejette H_0' au profit de H_1' si $t > c_\alpha$, ou :
On rejette H_0'' au profit de H_1'' si $t < -c_\alpha$.
Les valeurs critiques c_α (ou $c_{\alpha/2}$) sont les quantiles $t_{1-\alpha}$ (ou $t_{1-\alpha/2}$) d'une variable t de STUDENT à $m + n - 2$ degrés de liberté (cf. sect. 1.1.8).

1.1.5 Mise en œuvre

Le calcul de la statistique n'offre pas de difficulté particulière. Elle suit, sous H_0, une loi de STUDENT à $m + n - 2$ degrés de liberté.

1.1.6 Compléments

On préfère le test de MANN-WHITNEY-WILCOXON (cf. sect. 1.2) si les distributions des observations ne peuvent pas être postulées gaussiennes, ou si l'on craint une certaine hétérogénéité des variances des deux populations (le test de MANN-WHITNEY-WILCOXON y est relativement robuste, surtout si les tailles d'échantillons ne sont pas trop différentes) ; ou bien le test de la médiane (cf. sect. 1.3) si ces distributions sont très étalées (par exemple si la fiabilité des mesures est douteuse et si l'on peut s'attendre à une proportion peut-être faible mais non négligeable d'observations aberrantes).

1.1.7 Exemple

Dans une usine, la machine utilisée pour remplir des pots de yaourt se dérègle au cours du temps. En comparant les poids de 10 pots prélevés après cinq heures de fonctionnement (5 h) à ceux de 10 pots prélevés au début du fonctionnement de la machine (0 h), peut-on dire s'il faut recaler la machine?

0 h :	125,1	124,8	126,1	125,6	125,8	124,2	125,3	125,0	125,1	124,2
5 h :	128,2	127,9	129,4	128,8	128,2	128,0	128,4	128,1	129,1	127,9

Comme rien ne laisse prévoir le sens du dérèglement attendu, on fait un test bilatéral. On calcule facilement les quantités $m = n = 10$, $\overline{X} = 125,12$, $\sum(X_i - \overline{X})^2 = 3,496$, $\overline{Y} = 128,4$, $\sum(Y_i - \overline{Y})^2 = 2,48$, qui permettent de calculer

$$t = \frac{125,1 - 128,4}{\sqrt{(3,496 + 2,48)/18}\,\sqrt{(1/10) + (1/10)}} = -12,729.$$

On trouve dans la table, pour un risque $\alpha = 0,05 = 5\,\%$ et donc dans la colonne $\alpha/2 = 0,025$, avec $10 + 10 - 2 = 18$ degrés de liberté (ddl), une valeur critique $c_{\alpha/2} = 2,101$. Comme la statistique $|t|$ est très supérieure à $c_{\alpha/2}$, on rejette H_0 et on conclut qu'après cinq heures de fonctionnement, μ_Y est différent de μ_X : il faut donc recaler la machine.

1.1.8 Table

Pour n et α donnés, on trouve les valeurs critiques c_α ou $c_{\alpha/2}$ sur la ligne $ddl = m + n - 2$ de la table de quantiles des distributions t de STUDENT (sect. 1.1.8, p. 5).

1.1.9 Mise en œuvre informatique

S-PLUS Utiliser la fonctions `t.test`.

```
> yaourts.0<-c(125.1,124.8,126.1,125.6,125.8,124.2,125.3,
125.0,125.1,124.2)
> yaourts.5<-c(128.2,127.9,129.4,128.8,128.2,128.0,128.4,
128.1,129.1,127.9)
> res<-t.test(yaourts.0,yaourts.5,alternative="two.sided",
paired=F)
> summary(res)

Test de Student, deux echantillons non apparies :

 test de l'hypothese nulle :
 *H0 : "les moyennes sont egales",
 contre l'hypothese alternative bilaterale :
 *H1 : "les moyennes sont differentes".

 Conditions d'utilisation :
   les observations du premier echantillon sont independantes
   (tirage au sort des unites experimentales),
   les observations du deuxieme echantillon sont independantes
   (tirage au sort des unites experimentales),
   les deux echantillons sont independants
   la distribution des variables aleatoires est normale,
   les dispersions sont identiques.

------------------------------------------------------------
         Standard Two-Sample t-Test

data:  yaourts.0 and yaourts.5
t = -12.7289, df = 18, p-value = 0
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -3.82137 -2.73863
sample estimates:
 mean of x mean of y
```

```
    125.12     128.4

------------------------------------------------------------
Le test est realise au niveau de confiance 95%.
On rejette l'hypothese nulle, le risque d'erreur est egal a 0.
```

SAS procédure TTEST

```
data lost;
     input yaourt $ poids @@;
    cards;
y0 125.1 y0 124.8 y0 126.1 y0 125.6 y0 125.8
y0 124.2 y0 125.3 y0 125.0 y0 125.1 y0 124.2
y5 128.2 y5 127.9 y5 129.4 y5 128.8 y5 128.2
y5 128.0 y5 128.4 y5 128.1 y5 129.1 y5 127.9
;
proc ttest;
     class yaourt;
     var poids;
run;
```

```
                              TTEST PROCEDURE

Variable: POIDS

YAOURT     N     Mean Std Dev    Std Error Minimum Maximum
------------------------------------------------------------
y0        10  125.120   0.623   0.19708994  124.20  126.10
y5        10  128.400   0.525   0.16599866  127.90  129.40

Variances        T       DF    Prob>|T|
---------------------------------------
Unequal   -12.7289     17.5      0.0001
Equal     -12.7289     18.0      0.0000

For H0: Variances are equal, F' = 1.41    DF = (9,9)
Prob>F' = 0.6172
```

La procédure TTEST calcule la statistique de Student dans le cas où les variances sont considérées inégales (ligne Unequal) et dans le cas où elles sont considérées égales (ligne Equal). Le choix peut être guidé par la statistique F' qui permet de tester l'hypothèse de l'égalité des variances, cf. section 2.1.

1.2 Test de Mann-Whitney-Wilcoxon

Synonyme : test de la somme des rangs de WILCOXON.

1.2.1 Définition du problème : hypothèses nulle et alternative

Tester l'égalité des médianes m_X et m_Y des populations d'où sont tirés les deux échantillons $\{X_i,\, i = 1, \ldots, m\}$ et $\{Y_j,\, j = 1, \ldots, n\}$:
$H_0 \;:\; m_X = m_Y$ contre $H_1 \;:\; m_X \neq m_Y$ (alternative bilatérale), ou bien :
$H'_0 \;:\; m_X \leqslant m_Y$ contre $H'_1 \;:\; m_X > m_Y$ (alternative unilatérale), ou :
$H''_0 \;:\; m_X \geqslant m_Y$ contre $H''_1 \;:\; m_X < m_Y$ (alternative unilatérale).

1.2.2 Conditions d'utilisation

Les variables X_i et Y_j des deux échantillons doivent être continues pour éliminer la possibilité d'observations ex æquo, mais la procédure est relativement peu sensible à un nombre modéré d'ex æquo et le test s'applique donc aussi à des variables discrètes à valeurs ordonnées.

Les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent impérativement être indépendantes. En outre, les distributions des variables X_i et Y_j doivent être identiques à un éventuel décalage de médiane près.

1.2.3 Définition de la statistique, justification intuitive

On peut utiliser deux formes qui se déduisent facilement l'une de l'autre, le U de MANN et WHITNEY ou le S de WILCOXON (somme des rangs) :

- MANN et WHITNEY : on calcule U_X et U_Y :
  U_X = nombre de couples (X_i, Y_j) où X_i est supérieur à Y_j, et
  U_Y = nombre de couples (X_i, Y_j) où X_i est inférieur à Y_j ;
  (on remarque que $U_X + U_Y = mn$, nombre total de comparaisons) ;

- WILCOXON : on range par ordre croissant l'ensemble des $m+n$ valeurs des X_i et Y_j auxquelles on affecte ainsi leurs rangs, et on calcule :
  S_X = somme des rangs des valeurs de l'échantillon des X_i, et
  S_Y = somme des rangs des valeurs de l'échantillon des Y_j ;
  (on remarque que $S_X + S_Y = (m+n)(m+n+1)/2$, somme de tous les rangs de 1 à $m+n$) ;

On peut déduire les formes U des formes S (et réciproquement) en remarquant que $U_X = S_X - m(m+1)/2$ (où $m(m+1)/2$ est la plus petite valeur possible de S_X), et de même $U_Y = S_Y - n(n+1)/2$.

Intuitivement, on sent que les distributions des deux U seront, sous H_0, identiques et symétriques autour de leur moyenne commune $mn/2$; tandis que sous les hypothèses alternatives, elles seront dissymétriques et leurs médianes auront tendance à s'approcher du minimum 0 ou du maximum mn. On rejettera donc H_0 lorsque l'un des U s'écartera « trop » de $mn/2$.

1.2.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $U = \inf(U_X, U_Y) < U_\alpha$ (table bilatérale).
On rejette H_0' au profit de H_1' si $U_Y < U_\alpha$ (table unilatérale).
On rejette H_0'' au profit de H_1'' si $U_X < U_\alpha$ (table unilatérale).

1.2.5 Mise en œuvre

Calcul de la statistique
Compte tenu des relations entre elles, il suffit de déterminer l'une des formes rencontrées en section 1.2.3 pour en déduire facilement toutes les autres.

Distribution sous H_0
Pour m et $n \leqslant 20$, la distribution (commune sous H_0) de U_X ou U_Y est tabulée en section 1.2.8.
Pour m et n suffisamment grands (> 20), on transforme U_X ou U_Y, alors approximativement gaussiennes, en $Z = (U - mn/2)/\sqrt{mn(m+n+1)/12}$ et on utilise une table de la distribution gaussienne centrée réduite $\mathcal{N}(0,1)$ (ou la ligne « ∞ » de la table de la loi de STUDENT, cf. sect. 1.1.8, p. 5).

Ex æquo
S'il y a des ex æquo en quantité modérée, on calcule les statistiques S en attribuant des rangs moyens aux observations en cause (approche de WILCOXON), ou une contribution à U_X ou U_Y de seulement $1/2$ (et non 1 ou 0) pour chaque couple (X_i, Y_j) où $X_i = Y_j$ (approche de MANN et WHITNEY) ; on utilise alors les tables comme s'il n'y avait pas d'ex æquo (la procédure est conservative, c.-à-d. que le niveau du test reste contrôlé mais que l'on perd éventuellement un peu de puissance).
S'il y en a beaucoup, il faut soit recalculer la distribution de permutation de la statistique (seule possibilité avec de petits échantillons, malaisée sans un programme de calcul écrit à cette fin), soit apporter une correction au dénominateur de l'approximation pour grands échantillons (cf. SPRENT, 1992).

1.2.6 Compléments

Si de plus la distribution des X_i et Y_j peut être postulée gaussienne, le test t de STUDENT (cf. sect. 1.1) est préférable car légèrement plus puissant (asymptotiquement, il se contentera d'environ 5 % moins d'observations pour détecter une même différence significative). Par contre si cette distribution est très étalée (risque non négligeable de présence d'observations aberrantes), le test de la médiane (cf. sect. 1.3) pourrait être préférable.

1.2.7 Exemple

(*Tiré de* SPRENT, 1992).

Le problème : dans la bibliothèque du Pr. SPRENT, on regarde le nombre de pages de 12 livres de statistiques (rangés à part) et 16 autres livres divers. Les nombres médians de pages des populations de livres d'où sont tirés ces deux échantillons sont-ils statistiquement différents?
Nombres de pages des livres divers :
29 39 60 78 82 112 125 170 192 224 263 275 276 286 369 756
Nombres de pages des livres de statistiques :
126 142 156 228 245 246 370 419 433 454 478 503

Réalisation : ici, les tailles des échantillons sont $m = 16$ et $n = 12$. On range ensemble les $16 + 12 = 28$ valeurs des deux échantillons. À chacune est affecté son rang, comme dans le tableau suivant où l'on a souligné les valeurs et les rangs de l'échantillon de livres de statistiques :

Valeurs	29	39	60	78	82	112	125	<u>126</u>	<u>142</u>	<u>156</u>
Rangs	1	2	3	4	5	6	7	<u>8</u>	<u>9</u>	<u>10</u>
Valeurs	170	192	224	<u>228</u>	<u>245</u>	<u>246</u>	263	275	276	286
Rangs	11	12	13	<u>14</u>	<u>15</u>	<u>16</u>	17	18	19	20
Valeurs	369	<u>370</u>	<u>419</u>	<u>433</u>	<u>454</u>	<u>478</u>	<u>503</u>	756		
Rangs	21	<u>22</u>	<u>23</u>	<u>24</u>	<u>25</u>	<u>26</u>	<u>27</u>	28		

On calcule facilement la somme des rangs S_X des livres divers: $S_X = 1+2+3+4+5+6+7+11+12+13+17+18+19+20+21+28 = 187$, et donc $S_Y = (16+12)(16+12+1)/2 - 187 = 219$, $U_X = 187 - 16\times 17/2 = 51$, $U_Y = 16 \times 12 - 51 = 219 - 12 \times 13/2 = 141$ et $U = \inf(U_X, U_Y) = 51$. La valeur critique lue dans la table du test bilatéral de niveau $\alpha = 5\,\%$ avec $m = 16$ et $n = 12$ est $U_\alpha = 53$.

Comme $U < U_\alpha$, on rejette l'hypothèse d'égalité des médianes.

1.2.8 Tables de valeurs critiques du U de Mann et Whitney

Les tables ci-dessous donnent les valeurs critiques U_α : prendre la valeur à droite de la diagonale (|) pour le niveau 5 %, à gauche pour 1 %.

Test unilatéral

m \ n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	4|1	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6	2	7|3	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7	3	4	11|6	13	15	17	19	21	24	26	28	30	33	35	37	39
8	4	6	7	15|9	18	20	23	26	28	31	33	36	39	41	44	47
9	5	7	9	11	21|14	24	27	30	33	36	39	42	45	48	51	54
10	6	8	11	13	16	27|19	31	34	37	41	44	48	51	55	58	62
11	7	9	12	15	18	22	34|25	38	42	46	50	54	57	61	65	69

suite à la page suivante ...

... suite

m \ n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
12	8	11	14	17	21	24	28	42|31	47	51	55	60	64	68	72	77
13	9	12	16	20	23	27	31	35	51|39	56	61	65	70	75	80	84
14	10	13	17	22	26	30	34	38	43	61|47	66	71	77	82	87	92
15	11	15	19	24	28	33	37	42	47	51	72|56	77	83	88	94	100
16	12	16	21	26	31	36	41	46	51	56	61	83|66	89	95	101	107
17	13	18	23	28	33	38	44	49	55	60	66	71	96|77	102	109	115
18	14	19	24	30	36	41	47	53	59	65	70	76	82	109|88	116	123
19	15	20	26	32	38	44	50	56	63	69	75	82	88	94	123|101	130
20	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	138|104

Test bilatéral

m \ n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	2|0	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	1	5|2	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	1	3	8|4	10	12	14	16	18	20	22	24	26	28	30	32	34
8	2	4	6	13|7	15	17	19	22	24	26	29	31	34	36	38	41
9	3	5	7	9	17|11	20	23	26	28	31	34	37	39	42	45	48
10	4	6	9	11	13	23|16	26	29	33	36	39	42	45	48	52	55
11	5	7	10	13	16	18	30|21	33	37	40	44	47	51	55	58	62
12	6	9	12	15	18	21	24	37|27	41	45	49	53	57	61	65	69
13	7	10	13	17	20	24	27	31	45|34	50	54	59	63	67	72	76
14	7	11	15	18	22	26	30	34	38	55|42	59	64	69	74	78	83
15	8	12	16	20	24	29	33	37	42	46	64|51	70	75	80	85	90
16	9	13	18	22	27	31	36	41	45	50	55	75|60	81	86	92	98
17	10	15	19	24	29	34	39	44	49	54	60	65	87|70	93	99	105
18	11	16	21	26	31	37	42	47	53	58	64	70	75	99|81	106	112
19	12	17	22	28	33	39	45	51	57	63	69	74	81	87	113|93	119
20	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	127|105

1.2.9 Mise en œuvre informatique

S-PLUS Utiliser la fonction `wilcox.test`

```
> liv.stat<-c(126,142,156,228,245,246,370,419,433,454,478,503)
> romans<-c(29,39,60,78,82,112,125,170,192,224,263,275,276,
286,369,756)
> res<-wilcox.test(romans,liv.stat,alternative="two.sided",
paired=F)
> summary(res)

Test de Wilcoxon, deux echantillons non apparies :

 test de l'hypothese nulle :
```

```
 *H0 : "les medianes sont egales",
 contre l'hypothese alternative bilaterale :
 *H1 : "les medianes sont differentes".

 Conditions d'utilisation :
   les observations du premier echantillon sont independantes
   (tirage au sort des unites experimentales),
   les observations du deuxieme echantillon sont independantes
   (tirage au sort des unites experimentales),
   les deux echantillons sont independants
   les distributions des variables aleatoires sont quelconques
   elles ne different que par un parametre de position.

------------------------------------------------------------
         Exact Wilcoxon rank-sum test

data:  romans and liv.stat
rank-sum statistic W = 187, n = 16, m = 12, p-value = 0.0373
alternative hypothesis: true mu is not equal to 0

------------------------------------------------------------
Le test est realise au niveau de confiance 95%.
On rejette l'hypothese nulle, le risque d'erreur est egal
a 0.0373.
```

SAS procédure NPAR1WAY option wilcoxon

```
data lost;
     input livre $ npages @@;
    cards;
R 29 R 39 R 60 R 78 R 82 R 112 R 125 R 170 R 192
R 224 R 263 R 275 R 276 R 286 R 369 R 756
S 126 S 142 S 156 S 228 S 245 S 246 S 370
S 419 S 433 S 454 S 478 S 503
;
proc npar1way wilcoxon;
     class livre;
     var npages;
run;
```

```
                  N P A R 1 W A Y  P R O C E D U R E

            Wilcoxon Scores (Rank Sums) for Variable NPAGES
                  Classified by Variable LIVRE


                   Sum of     Expected      Std Dev      Mean
 LIVRE      N     Scores     Under H0     Under H0     Score

 R         16      187.0        232.0   21.5406592   11.6875
 S         12      219.0        174.0   21.5406592   18.2500

       Wilcoxon 2-Sample Test (Normal Approximation)
```

```
(with Continuity Correction of .5)

S=  219.000      Z=  2.06586      Prob > |Z| =   0.0388

T-Test approx. Significance =     0.0486

Kruskal-Wallis Test (Chi-Square Approximation)
CHISQ=  4.3642      DF=  1      Prob > CHISQ=    0.0367
```

On rejette l'hypothèse nulle avec le risque de 0.0388

1.3 Test de la médiane

1.3.1 Définition du problème: hypothèses nulle et alternative

Tester l'égalité des médianes m_X et m_Y des populations d'où sont tirés les deux échantillons $\{X_i,\, i = 1, \ldots, m\}$ et $\{Y_j,\, j = 1, \ldots, n\}$:
$H_0 \ : \ m_X = m_Y$ contre $H_1 \ : \ m_X \neq m_Y$ (alternative bilatérale), ou bien :
$H_0' \ : \ m_X \leqslant m_Y$ contre $H_1' \ : \ m_X > m_Y$ (alternative unilatérale), ou :
$H_0'' \ : \ m_X \geqslant m_Y$ contre $H_1'' \ : \ m_X < m_Y$ (alternative unilatérale).

1.3.2 Conditions d'utilisation

Les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent impérativement être indépendantes, et au moins ordinales. En outre, si les médianes sont égales, la probabilité de dépasser la médiane commune doit être la même dans les deux populations (et égale à $1/2$ si la probabilité d'une valeur exactement égale à cette médiane est nulle, par exemple si les distributions sont continues).

1.3.3 Définition de la statistique, justification intuitive

On détermine la médiane $\widetilde{m}_{X,Y}$ de l'ensemble des $m + n$ valeurs. On peut alors ranger les nombres d'observations respectivement supérieures et inférieures à $\widetilde{m}_{X,Y}$ de chacun des deux échantillons dans une table 2×2 :

	$X_1, \ldots, X_m$	$Y_1, \ldots, Y_n$	Totaux
$> \widetilde{m}_{X,Y}$	x	y	$a = x + y \ (\approx \frac{m+n}{2})$
$< \widetilde{m}_{X,Y}$	u	v	$b = u + v \ (\approx \frac{m+n}{2})$
Totaux	$c = x + u$ $(\approx m)$	$d = y + v$ $(\approx n)$	$t = a + b = c + d$ $(\approx m + n)$

On utilise le nombre x de valeurs du premier échantillon strictement supérieures à $\widetilde{m}_{X,Y}$. En ignorant l'éventualité de valeurs égales à $\widetilde{m}_{X,Y}$, la

moyenne de la variable x est $m/2$ sous H_0 tandis que sous les alternatives elle aura tendance à s'en éloigner (par valeurs supérieures si la médiane des X_i dépasse celle des Y_j, ou inférieures sinon). La différence de symétrie dans le traitement des deux échantillons n'est qu'apparente: x et y apportent la *même* information, car on peut considérer que le nombre total $(x+y)$ d'observations supérieures à $\widetilde{m}_{X,Y}$ est *fixé* à $(m+n)/2$. De même, u et v apportent aussi la même information ($x+u=m$ et $y+v=n$).

1.3.4 Régions de rejet et de non-rejet

Dans le cas de petits échantillons, on utilise le fait que sous H_0, la statistique x suit une loi hypergéométrique (loi du nombre de boules blanches tirées) de paramètres t (nombre de boules dans l'urne), c (nombre de boules blanches dans l'urne), a (nombre de boules tirées). Dans le cas unilatéral, on calcule la P-variable associée à la valeur x observée, et on rejette H_0 si elle est inférieure au niveau choisi (cf. ex. en sect. 1.3.7). Dans le cas bilatéral, la notion de P-variable est un peu délicate à définir car la distribution de la statistique (distribution hypergéométrique) n'est pas *symétrique* en général. La construction de la région de rejet peut se faire de la manière suivante. On ordonne les réalisations possibles de la statistique par probabilités croissantes. La région de rejet au niveau α est l'ensemble des premières valeurs dont la somme des probabilités n'excède pas α.

Dans le cas de grands échantillons (à partir de m et n de l'ordre d'une quinzaine), on utilise une approximation normale de la loi hypergéométrique. On définit $X^* = x+1/2$ si $x < m/2$ et $X^* = x-1/2$ si $x > m/2$ (correction de continuité), et la variable $Z = (X^* - m/2)/\sqrt{mn/4(m+n)}$ suit alors sous H_0 une loi approximativement gaussienne centrée réduite. Notons z_α le quantile d'ordre α de cette distribution $\mathcal{N}(0,1)$.

On rejette H_0 au profit de H_1 si $|Z| > z_{1-\alpha/2}$.
On rejette H_0' au profit de H_1' si $Z > z_{1-\alpha}$.
On rejette H_0'' au profit de H_1'' si $Z < z_\alpha = -z_{1-\alpha}$.

1.3.5 Mise en œuvre

Le calcul de la statistique ne présente pas de difficulté particulière.

Les ex æquo ne posent problème que s'ils sont égaux à la médiane $\widetilde{m}_{X,Y}$: la moyenne et la variance asymptotiques de X^* sont modifiés. S'il n'y en a pas trop, on peut les éliminer et réduire m et n en conséquence.

1.3.6 Compléments

Ce test est un cas particulier du *test exact de* FISHER pour les tables 2×2 (cf. SPRENT 1992) : la distribution de x sous H_0 est hypergéométrique (malpratique sauf avec un programme de calcul adapté, c'est pourquoi nous ne traitons le cas des petits échantillons que dans l'exemple, cf. sect. 1.3.7).

Si leurs conditions d'utilisation (plus strictes) sont réunies, on préfère les tests (alors plus puissants) de MANN-WHITNEY-WILCOXON ou de STUDENT.

1.3.7 Exemple

(*Tiré de* SPRENT, 1992).

Reprenons l'exemple de la p. 54 : les nombres médians de pages des livres de statistiques et des livres « divers » de la bibliothèque du Pr. SPRENT, d'où sont tirés les échantillons suivants, sont-ils statistiquement différents ?

Nombres de pages de 16 livres « divers » :
29 39 60 78 82 112 125 170 192 224 263 275 276 286 369 756
Nombres de pages de 12 livres de statistiques :
126 142 156 228 245 246 370 419 433 454 478 503

On voit facilement que seuls 6 des 16 livres « divers » sont supérieurs à la médiane $(228+245)/2$ de l'ensemble des observations, d'où la table 2×2 :

	Livres divers	Livres de statistiques	Totaux
$> \widetilde{m}_{X,Y}$	$x=6$	$y=8$	$a=14\ (\equiv \frac{m+n}{2})$
$< \widetilde{m}_{X,Y}$	$u=10$	$v=4$	$b=14\ (\equiv \frac{m+n}{2})$
Totaux	$c=16\ (\equiv m)$	$d=12\ (\equiv n)$	$t=28\ (\equiv m+n)$

La distribution de la variable x est celle du nombre de boules blanches dans un échantillon de taille 14 tiré (sans remise) d'une urne lorsque celle-ci contient 16 boules blanches sur un total de 28. Il s'agit d'une distribution hypergéométrique. Les probabilités correspondantes sont données dans le tableau suivant :

x	2	3	4	5	6	7	8
Pr	0,000	0,000	0,003	0,024	0,099	0,226	0,296
x	9	10	11	12	13	14	
Pr	0,226	0,099	0,024	0,003	0,000	0,000	

On observera que la distribution de x est symétrique. Cette symétrie est due aux valeurs particulières de a, c et t (si on avait $t=27$ au lieu de $t=28$, la distribution ne serait pas symétrique). On peut donc calculer la P-variable associée à $x=6$ et on trouve

$$2\times(0,000+0,000+0,003+0,024+0,099)=0,25.$$

Ce test ne rejette donc pas au niveau 5 % l'hypothèse d'égalité des médianes (différence de puissance des deux tests).

La statistique de l'approximation gaussienne pour grands échantillons vaut $Z=(6+1/2-16/2)/\sqrt{16\times 12/[4\times(16+12)]}=-1,146$, inférieure (en valeur absolue) à la valeur critique 1,96 du test bilatéral de niveau 5 % : on ne rejette toujours pas l'hypothèse de médianes égales.

1.3.8 Table

On trouve les z_α dans les tables de la loi gaussienne standard $\mathcal{N}(0,1)$, ou sur la ligne « ∞ » de la table de la loi de STUDENT, cf. sect. 1.1.8, p. 5.

1.3.9 Mise en œuvre informatique

S-PLUS Utiliser la fonction `NESI` de la bibliothèque du même nom.

```
> library(NESI)
> liv.stat<-c(126,142,156,228,245,246,370,419,433,454,478,503)
> romans<-c(29,39,60,78,82,112,125,170,192,224,263,275,276,
286,369,756)
> res<-NESI(romans,liv.stat)

"median.test(...)"
> summary(res)

Test de la mediane, deux echantillons non apparies :

 test de l'hypothese nulle :
 *H0 : "les medianes sont egales",
 contre l'hypothese alternative bilaterale :
 *H1 : "les medianes sont differentes".

 Conditions d'utilisation :
  les observations du premier echantillon sont independantes
  (tirage au sort des unites experimentales),
  les observations du deuxieme echantillon sont independantes
  (tirage au sort des unites experimentales),
  les deux echantillons sont independants,
  les distributions des variables aleatoires sont quelconques,
  elles peuvent differer par leur forme et (ou) par leur
  dispersion.

------------------------------------------------------------
         Exact Median Test

data:  romans and liv.stat
statistic = 6, m = 16, n = 12, p-value = 0.2519
alternative hypothesis: true mu is not equal to 0
sample estimates:
 median of x median of y
         181         308

------------------------------------------------------------
Le test est realise au niveau de confiance 95%.
On ne rejette pas l'hypothese nulle, on accepte l'egalite
des medianes.
```

NESI a utilisé la loi hypergéométrique pour calculer la P-variable. La méthode de calcul de celle-ci pour un test bilatéral est expliquée en section 1.3.4.

SAS Procédure NPAR1WAY option median

```
data lost;
     input livre $ npages @@;
    cards;
R 29 R 39 R 60 R 78 R 82 R 112 R 125 R 170 R 192
R 224 R 263 R 275 R 276 R 286 R 369 R 756
S 126 S 142 S 156 S 228 S 245 S 246 S 370
S 419 S 433 S 454 S 478 S 503
;
proc npar1way median;
     class livre;
     var npages;
run;
```

```
                N P A R 1 W A Y  P R O C E D U R E

        Median Scores (Number of Points above Median)
                    for Variable NPAGES
                  Classified by Variable LIVRE


                Sum of     Expected    Std Dev      Mean
  LIVRE    N   Scores     Under H0    Under H0     Score

  R       16      6.0          8.0      1.3333   0.37500
  S       12      8.0          6.0      1.3333   0.66667


        Median 2-Sample Test (Normal Approximation)
    S=  8.00000      Z=  1.50000      Prob > |Z| =   0.1336

        Median 1-Way Analysis (Chi-Square Approximation)
    CHISQ=  2.2500      DF=  1      Prob > CHISQ= 0.1336
```

SAS a utilisé une approximation normale pour la distribution de la statistique de test. La P-variable est donnée par `Prob >|Z|`. Comme $0,1336 > 0,05$, on ne rejette pas l'hypothèse nulle. On notera que la P-variable obtenue par approximation gaussienne est sensiblement différente de celle obtenue à partir de la loi hypergéométrique par S-PLUS.

La valeur de Z obtenue par SAS ($Z = 1,5$) n'est pas la même que celle obtenue dans l'exemple (section 1.3.7, $Z = -1.146$). Cette différence a des causes multiples. Premièrement Z est calculé à partir du nombre de livres de statistique dont le nombre de pages est inférieur à la médiane, alors que dans la section 1.3.7, on est parti du nombre de romans dont le nombre de pages est inférieur à la médiane. Deuxièmement, SAS ne fait pas de correction de continuité. Enfin l'approximation gaussienne utilisée par SAS est légèrement différente de celle donnée en section 1.3.4 et utilisée en section

1.3.7. Le calcul du Z par SAS utilise la formule suivante :

$$Z = \frac{y - na/t}{\sqrt{\frac{abcd}{t^2(t-1)}}}.$$

Chapitre 2

Comparaison des dispersions

2.1 Test d'égalité des variances F de Fisher-Snedecor

2.1.1 Définition du problème: hypothèses nulle et alternative

Tester l'égalité des variances σ_X^2 et σ_Y^2 des populations d'où sont tirés les deux échantillons $\{X_i,\, i = 1, \ldots, m\}$ et $\{Y_j,\, j = 1, \ldots, n\}$:

$H_0 \ : \ \sigma_X = \sigma_Y$ contre $H_1 \ : \ \sigma_X \neq \sigma_Y$ (alternative bilatérale), ou bien :
$H_0' \ : \ \sigma_X \leqslant \sigma_Y$ contre $H_1' \ : \ \sigma_X > \sigma_Y$ (alternative unilatérale), ou :
$H_0'' \ : \ \sigma_X \geqslant \sigma_Y$ contre $H_1'' \ : \ \sigma_X < \sigma_Y$ (alternative unilatérale).

2.1.2 Conditions d'utilisation

On doit avoir deux échantillons indépendants, c.-à-d. que les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent être indépendantes. De plus, les distributions des variables X_i et Y_j doivent impérativement être gaussiennes, de variances inconnues σ_X^2 et σ_Y^2 resp. (les moyennes n'intervenant pas).

2.1.3 Définition de la statistique, justification intuitive

$\widehat{\sigma}_X^2/\sigma_X^2 = \sum(X_i - \overline{X})^2/\sigma_X^2(m-1)$ et $\widehat{\sigma}_Y^2/\sigma_Y^2 = \sum(Y_j - \overline{Y})^2/\sigma_Y^2(n-1)$ sont des variables χ^2 divisées par leurs nombres de degrés de liberté ($m-1$ et $n-1$) et indépendantes. Par conséquent, dans tous les cas, leur rapport :

$$\frac{\sum_{i=1}^m (X_i - \overline{X})^2}{\sigma_X^2(m-1)} \Bigg/ \frac{\sum_{j=1}^n (Y_j - \overline{Y})^2}{\sigma_Y^2(n-1)} = \left(\frac{\widehat{\sigma}_X^2}{\widehat{\sigma}_Y^2}\right) \Bigg/ \left(\frac{\sigma_X^2}{\sigma_Y^2}\right) = F_0 \Bigg/ \left(\frac{\sigma_X^2}{\sigma_Y^2}\right)$$

suit une loi F de Fisher-Snedecor à $m-1$ et $n-1$ degrés de liberté. La statistique $F_0 = \widehat{\sigma}_X^2/\widehat{\sigma}_Y^2$ suit cette même loi *sous* H_0 (car alors $\sigma_X^2/\sigma_Y^2 = 1$) tandis que sous les alternatives, F_0 est la même variable *multipliée* par ce rapport, inconnu mais plus grand que 1 sous H_1' (resp. plus petit sous H_1''), et aura donc plus tendance à être « trop » grande (resp. trop petite).

2.1.4 Régions de rejet et de non-rejet

Soit $F_{1-\alpha,p,q}$ le quantile d'ordre $1-\alpha$ d'une variable F de FISHER-SNEDECOR à p et q degrés de liberté (valeur, que l'on trouve dans la table de la section 2.1.8, telle que $\Pr\{F < F_{1-\alpha,p,q}\} = 1-\alpha$). On rejette :
H_0 au profit de H_1 si $F_0 > F_{1-\alpha/2,m-1,n-1}$ ou si $F_0 < 1/F_{1-\alpha/2,n-1,m-1}$,
H'_0 au profit de H'_1 si $F_0 > F_{1-\alpha,m-1,n-1}$,
H''_0 au profit de H''_1 si $F_0 < 1/F_{1-\alpha,n-1,m-1}$.

2.1.5 Mise en œuvre

Les tables donnent seulement les quantiles d'ordre $1-\alpha$ et non ceux d'ordre α (nécessaires par exemple pour juger si F_0 est « trop petite ») car on peut déterminer les seconds à partir des premiers.

En effet, si F est une variable de FISHER-SNEDECOR à p et q degrés de liberté, $1/F$ est aussi, par définition, une variable de FISHER-SNEDECOR, mais à q et p degrés de liberté (remarquer l'interversion de p et q). Or $\Pr\{F < F_{1-\alpha,p,q}\} = \Pr\{1/F > 1/F_{1-\alpha,p,q}\} = 1 - \Pr\{1/F < 1/F_{1-\alpha,p,q}\}$ vaut $1-\alpha$ par définition de $F_{1-\alpha,p,q}$. Donc $\Pr\{1/F < 1/F_{1-\alpha,p,q}\} = \alpha$, ce qui signifie que le quantile d'ordre α de $1/F$ s'écrit $F_{\alpha,q,p} = 1/F_{1-\alpha,p,q}$, ou encore, en échangeant p et q, que le quantile d'ordre α de F (que l'on cherche) s'écrit $F_{\alpha,p,q} = 1/F_{1-\alpha,q,p}$; d'où la forme des régions de rejet.

2.1.6 Compléments

Ce test est très sensible à la condition de normalité des distributions. En cas de doute, on préférera le test de SIEGEL-TUKEY (cf. sect. 2.2).

2.1.7 Exemple

Un magazine féminin pense pouvoir affirmer qu'au-delà de 50 ans, les femmes continuent à suivre leur ligne, tandis que les hommes n'y prennent plus garde. On peut penser que si c'est le cas, la dispersion du poids des femmes sera inférieure à celle des hommes.

Voici les résultats (poids en kg) du tirage de deux échantillons ($m = 10$ et $n = 9$) effectué pour confirmer ou infirmer cette présomption :

Hommes (x_i) :	75	71,2	75,2	88,2	80	63	78	81	102	92
Femmes (y_j) :	62	50	49	53,5	55,2	57,1	54,8	62,4	65,2	

On calcule facilement $\widehat{\sigma}_X^2 = 123,509$, $\widehat{\sigma}_Y^2 = 31,642$ ainsi que leur rapport $F_0 = 3,903$. Dans la table, on lit $F_{95\,\%,9,8} = 3,39$. Comme $F_0 > F_{95\,\%,9,8}$, on rejette $H'_0 : \sigma_X \leqslant \sigma_Y$ au profit de l'alternative $H'_1 : \sigma_X > \sigma_Y$, c'est-à-dire que le sondage confirme l'intuition de la rédaction du magazine.

2.1.8 Tables de quantiles des lois de Fisher-Snedecor

Les tables ci-dessous donnent les quantiles $F_{1-\alpha,p,q}$ des lois de Fisher-Snedecor pour $\alpha = 5\%$ et $\alpha = 2,5\%$.

$\alpha = 5\,\%$

q \ p :	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

$\alpha = 5\,\%$

q	p : 12	15	20	24	30	40	60	120	∞
1	243,9	254,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

$\alpha = 2{,}5\,\%$

q	p : 1	2	3	4	5	6	7	8	9	10
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40

suite à la page suivante ...

α = 2,5 %

... suite

q \ p :	1	2	3	4	5	6	7	8	9	10
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,76	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,43
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67
24	5,72	4,32	3,72	3,28	3,15	2,99	2,87	2,78	2,70	2,64
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	1,27
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05

α = 2,5 %

q \ p :	12	15	20	24	30	40	60	120	∞
1	976,7	984,9	993,1	997,2	1001	1006	1010	1014	1018
2	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26

$\alpha = 2{,}5\,\%$

... suite

q \ p :	12	15	20	24	30	40	60	120	∞
5	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,92	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

2.1.9 Mise en œuvre informatique

S-PLUS Utiliser la fonction `var.test`.

```
> poids.hommes<-c(75,71,2,75.2,88.2,80.0,63.0,78.0,81.0,102.0,
92.0)
> poids.femmes<-c(62,50,49,53.5,55.2,57.1,54.8,62.4,65.2)
> res<-var.test(poids.hommes,poids.femmes,
alternative="greater")
> res
```

```
        F test for variance equality

data:  poids.hommes and poids.femmes
F = 3.9033, num df = 9, denom df = 8, p-value = 0.0341
alternative hypothesis: true ratio of variances is greater
than 1
95 percent confidence interval:
 1.152064       NA
sample estimates:
 variance of x variance of y
      123.5093      31.64194
```

La P-variable est inférieure à $0,05$, on conclut donc que la variance du poids des hommes est supérieur à la variance du poids des femmes.

SAS

```
data phf;
     input sexe $ poids @@;
    cards;
M 75.0 M 71.2 M 75.2 M 88.2 M 80.0 M 63.0 M 78.0 M 81.0
M 102.0 M 92.0
F 62.0 F 50.0 F 49.0 F 53.5 F 55.2 F 57.1 F 54.8 F 62.4 F 65.2
;

proc ttest;
     class sexe;
     var poids;
run;
```

```
                              TTEST PROCEDURE

Variable: POIDS

SEXE    N    Mean      Std Dev    Std Error Minimum  Maximum
-------------------------------------------------------------
F       9  56.578   5.62511728   1.87503909  49.000   65.200
M      10  80.560  11.11347530   3.51438947  63.000  102.000

Variances        T       DF    Prob>|T|
----------------------------------------
Unequal    -6.0207     13.6      0.0001
Equal      -5.8255     17.0      0.0000

For H0: Variances are equal,
F' = 3.90  DF = (9,8)  Prob>F' = 0.0682
```

Pour avoir la probabilité correspondant à l'hypothèse H_1 unilatérale, il suffit de diviser par 2 Prob>F', on obtiend donc 0.0341.

2.2 Test d'égalité des dispersions de Siegel et Tukey

2.2.1 Définition du problème: hypothèses nulle et alternative

Tester l'égalité des paramètres de dispersion σ_X^2 et σ_Y^2 (par exemple les variances) des populations d'où sont tirés les deux échantillons observés :
$H_0 : \sigma_X = \sigma_Y$ contre $H_1 : \sigma_X \neq \sigma_Y$ (alternative bilatérale), ou bien :
$H_0' : \sigma_X \leqslant \sigma_Y$ contre $H_1' : \sigma_X > \sigma_Y$ (alternative unilatérale), ou :
$H_0'' : \sigma_X \geqslant \sigma_Y$ contre $H_1'' : \sigma_X < \sigma_Y$ (alternative unilatérale).

2.2.2 Conditions d'utilisation

Les variables X_i et Y_j des deux échantillons doivent être au moins ordinales. On doit avoir deux échantillons indépendants, c.-à-d. que les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent impérativement être indépendantes.

De plus, les distributions des variables $(X_i - \mu)/\sigma_X$ et $(Y_j - \mu)/\sigma_Y$, où μ est un paramètre de position (moyenne ou médiane par exemple) commun aux deux populations, doivent être identiques. Il n'y a donc pas de postulat de normalité, contrairement au test de Fisher-Snedecor.

2.2.3 Définition de la statistique, justification intuitive

La statistique utilisée ressemble beaucoup à celle du test de Mann-Whitney-Wilcoxon (et sa distribution sous H_0 est la même) : on réunit les deux échantillons et on attribue à chaque observation un des entiers compris entre 1 et $m+n$. Cependant, on n'associe pas les plus petits aux valeurs les plus petites (comme avec les rangs) mais aux valeurs les plus éloignées de la médiane générale (voir détails en sect. 2.2.5) : si les dispersions diffèrent, la somme des entiers (ou la forme U équivalente) associée à l'un des échantillons (le plus dispersé) a plus tendance que sous H_0 à être « trop » faible.

2.2.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $U = \inf(U_X, U_Y) < U_\alpha$ (table bilatérale).
On rejette H_0' au profit de H_1' si $U_Y < U_\alpha$ (table unilatérale).
On rejette H_0'' au profit de H_1'' si $U_X < U_\alpha$ (table unilatérale).

2.2.5 Mise en œuvre

Pour le calcul de la statistique, on attribue aux $m+n$ valeurs rangées par ordre croissant l'entier 1 à la plus petite, puis 2 à la plus grande et 3 à la deuxième plus grande, 4 et 5 aux deuxième et troisième plus petites, 6 et 7 aux deux plus grandes suivantes, et ainsi de suite (cf. p. ex. sect. 2.1.7) jusqu'à $(m+n)$. Puis on calcule les sommes S_X et S_Y des entiers attribués

aux X_i et Y_j (resp.) ainsi que $U_X = S_X - m(m+1)/2$ et $U_Y = S_Y - n(n+1)/2$ (on a encore $S_X + S_Y = (m+n)(m+n+1)/2$ et $U_X + U_Y = mn$).

2.2.6 Compléments

Lorsqu'une différence de position ne peut être exclue, on peut ajouter une estimation de cette différence à chaque valeur de l'échantillon dont la position est plus faible. Mais le test sera alors conditionnel à cette estimation, et la distribution de la table ne sera plus en toute rigueur applicable : on ne peut conseiller cette procédure qu'avec des échantillons pas trop petits.

2.2.7 Exemple

(*Tiré de* SPRENT, 1992).

Reprenons l'exemple de la p. 54 : les dispersions des nombres de pages des livres de statistiques et des livres « divers » de la bibliothèque du Pr. SPRENT d'où sont tirés les échantillons suivants sont-elles statistiquement différentes ?

Nombres de pages de $m = 16$ livres « divers » :

29 39 60 78 82 112 125 170 192 224 263 275 276 286 369 756

Nombres de pages de $n = 12$ livres de statistiques :

126 142 156 228 245 246 370 419 433 454 478 503

Nous avions conclu à l'existence d'une différence entre les médianes, que l'on peut estimer par la médiane 133,5 des mn différences $Y_j - X_i$. Si l'on ajoute cette valeur aux nombres de pages des livres divers (non soulignés) et si l'on refait le rangement par ordre croissant, on obtient, en attribuant les entiers de 1 à 28 selon la méthode décrite plus haut, le tableau :

Valeurs	126	142	156	162,5	172,5	193,5	211,5	215,5	228	245
Entiers	1	4	5	8	9	12	13	16	17	20
Valeurs	245,5	246	258,5	303,5	325,5	357,5	370	396,5	408,5	409,5
Entiers	21	24	25	28	27	26	23	22	19	18
Valeurs	419	419,5	433	454	478	502,5	503	889,5		
Entiers	15	14	11	10	7	6	3	2		

La somme S_Y des 12 livres de statistiques vaut 140 ; donc $U_Y = 62$ et $U_X = 130$, et $U = 62$ est supérieure à la valeur critique $U_{5\,\%} = 53$ du test bilatéral : on ne peut donc rejeter l'hypothèse d'égalité des dispersions.

2.2.8 Table

La distribution de la statistique est identique à celle de la forme U du test de MANN-WHITNEY-WILCOXON : on se reportera à sa table (sect. 1.2.8).

Chapitre 3

Comparaison globale des répartitions

3.1 Test χ^2 d'homogénéité de deux populations

3.1.1 Définition du problème: hypothèses nulle et alternative

Tester l'égalité des répartitions d'une variable discrète (c valeurs possibles) ou comparer les fréquences de c catégories dans deux populations X et Y. En notant $p_{X,i}$ et $p_{Y,i}$ les probabilités de la catégorie i dans les populations X et Y, on teste donc H_0 : $p_{X,i} = p_{Y,i}$ pour chacune des classes $(i = 1, \ldots, c)$, contre H_1 : il existe au moins une classe i telle que $p_{X,i} \neq p_{Y,i}$.

3.1.2 Conditions d'utilisation

Les données se présentent sous forme d'une table de contingence $2 \times c$ rassemblant les effectifs observés dans les diverses classes :

	Classe C_1	Classe C_2	...	Classe C_c	Totaux
Population X	$n_{X,1}$	$n_{X,2}$	...	$n_{X,c}$	N_X
Population Y	$n_{Y,1}$	$n_{Y,2}$	...	$n_{Y,c}$	N_Y
Totaux	N_1	N_2	...	N_c	N

Les classes peuvent être des catégories sans ordre entre elles (variable discrète nominale), et c'est le cadre naturel d'utilisation de ce test. Mais il peut aussi y avoir un ordre (variable discrète ordinale, ou même variable continue à valeurs regroupées en classes) bien qu'on puisse alors préférer le test de Smirnov (cf. sect. 3.2), sauf dans le cas d'une variable discrète à très peu de valeurs possibles et avec des observations nombreuses.

3.1.3 Définition de la statistique, justification intuitive

Sous H_0, les probabilités $p_{X,i}$ et $p_{Y,i}$ sont égales mais cette valeur est inconnue : on doit l'estimer, par le rapport N_i/N, pour pouvoir estimer les effectifs espérés $e_{X,i} = N_X(N_i/N)$ et $e_{Y,i} = N_Y(N_i/N)$ dans la classe i pour les populations X et Y. La statistique utilisée est alors :

$$\chi^2 = \sum_{i=1}^{i=c} \left[\frac{(n_{X,i} - e_{X,i})^2}{e_{X,i}} + \frac{(n_{Y,i} - e_{Y,i})^2}{e_{Y,i}} \right] = N_X N_Y \sum_{i=1}^{i=c} \frac{1}{N_i} \left(\frac{n_{X,i}}{N_X} - \frac{n_{Y,i}}{N_Y} \right)^2$$

où la première forme est la classique somme, sur l'ensemble des cellules, de termes du type *(effectif observé – effectif espéré)² / effectif espéré*, tandis que la seconde, algébriquement équivalente, montre que l'on fait une somme des carrés (pondérés par l'inverse des effectifs des classes) des écarts entre les estimations des probabilités des classes pour chacune des deux populations.

3.1.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $\chi^2 \geqslant \chi^2_{\alpha,c-1}$, le $(1-\alpha)$-quantile de la distribution χ^2 que l'on trouve dans la table de la section 2.4.8 (p. 34).

3.1.5 Mise en œuvre

Le calcul de la statistique ne présente pas de particularité.

La distribution χ^2 sous H_0 n'est valide qu'asymptotiquement. En pratique on accepte l'approximation, lorsque $c = 2$, si tous les effectifs espérés $e_{X,i}$ et $e_{Y,i}$ sont supérieurs à 5, et lorsque $c > 2$, si plus de 75 % d'entre eux sont supérieurs à 5 et si aucun n'est inférieur à 1. Si ces conditions ne sont pas remplies, il faut soit regrouper des classes (si cela a un sens), soit réaliser plus d'observations, soit renoncer au test.

3.1.6 Compléments

Lorsque les classes sont ordonnées, les tests de Smirnov (cf. sect. 3.2) ou de Cramér-von Mises (cf. sect. 3.3) sont en général préférables si ces classes proviennent du regroupement de valeurs continues.

Ce test est un cas particulier du test χ^2 d'homogénéité pour trois populations ou plus.

3.1.7 Exemple

On demande à des élèves d'un collège leurs sports préférés. Leurs réponses sont présentées dans le tableau suivant.

	Athlétisme	Football	Volley-ball	Autres	Totaux
Filles	30	8	47	27	112
Garçons	21	81	18	14	134
Totaux	51	89	65	41	246

Peut-on, à partir de cet échantillon, considérer que les préférences se répartissent de la même manière chez les filles et les garçons ?

On commence par calculer le tableau des effectifs espérés. Par exemple, en notant X la population des filles, Y celle des garçons et en numérotant les classes dans l'ordre du tableau des données, $e_{X,1} = 51 \times 112/246 \approx 23,22$, $e_{Y,4} = 41 \times 134/246 \approx 22,33$. On obtient ainsi le tableau suivant :

	Athlétisme	Football	Volley-ball	Autres
Filles	23,22	40,52	29,59	18,67
Garçons	27,78	48,48	35,41	22,33

On en tire $\chi^2 = \frac{(30-23,22)^2}{23,22} + \frac{(21-27,78)^2}{27,78} + \cdots + \frac{(14-22,33)^2}{22,33} \approx 77,175$.

Par ailleurs, on trouve dans la table de quantiles des distributions χ^2 la valeur critique $\chi^2_{0,950,3} = 7,815$, bien inférieure à la valeur de la statistique : on rejette très clairement l'hypothèse de préférences identiques chez les filles et les garçons.

3.1.8 Table

Pour c et α donnés, on trouve les valeurs critiques $\chi^2_{\alpha,c-1}$ dans la table de quantiles des distributions χ^2 (cf. sect. 2.4.8, p. 34), sur la ligne $\nu = c-1$.

3.1.9 Mise en œuvre informatique

S-PLUS Utiliser la fonction `chisq.test`.

```
> sports<-matrix(c(30,21,8,81,47,18,27,14),nrow=2,ncol=4)
> res<-chisq.test(sports)
> res

  Pearson's chi-square test without Yates' continuity
                    correction

data:  sports
X-squared = 77.1748, df = 3, p-value = 0
```

SAS

```
data asc;
      do a=1 to 4;
         do b=1 to 2;
            input adept@@;
            output;
         end;
      end;
    cards;
30 21 8 81 47 18 27 14
;

proc freq;
     weight adept;
     tables a*b/chisq;
run;
```

TABLE OF A BY B

A	B		
Frequency |			
Percent |			
Row Pct |			
Col Pct |	1	2	Total
1	30	21	51
	12.20	8.54	20.73
	58.82	41.18	
	26.79	15.67	
2	8	81	89
	3.25	32.93	36.18
	8.99	91.01	
	7.14	60.45	
3	47	18	65
	19.11	7.32	26.42
	72.31	27.69	
	41.96	13.43	
4	27	14	41
	10.98	5.69	16.67
	65.85	34.15	
	24.11	10.45	
Total	112	134	246
	45.53	54.47	100.00

STATISTICS FOR TABLE OF A BY B

Statistic	DF	Value	Prob

```
-----------------------------------------------------
Chi-Square                      3    77.175       0.000
Likelihood Ratio Chi-Square     3    86.801       0.000
Mantel-Haenszel Chi-Square      1    12.334       0.000
Phi Coefficient                       0.560
Contingency Coefficient               0.489
Cramer's V                            0.560

Sample Size = 246
```

Les résultats du test du χ^2 sont donnés dans la ligne "Chi-Square...".

3.2 Test de Smirnov (ou Kolmogorov-Smirnov)

3.2.1 Définition du problème: hypothèses nulle et alternative

Tester l'identité des fonctions de répartition F_X et F_Y des populations d'où sont tirés les échantillons $\{X_i,\, i = 1, \ldots, m\}$ et $\{Y_j,\, j = 1, \ldots, n\}$:

$H_0 : F_X(z) = F_Y(z)$ pour tout réel z contre $H_1 : F_X(z) \neq F_Y(z)$ pour au moins un réel z (alternative bilatérale), ou bien :

$H_0' : F_X(z) \leqslant F_Y(z)$ pour tout réel z contre $H_1' : F_X(z) > F_Y(z)$ pour au moins un réel z (alternative unilatérale), ou :

$H_0'' : F_X(z) \geqslant F_Y(z)$ pour tout réel z contre $H_1'' : F_X(z) < F_Y(z)$ pour au moins un réel z (alternative unilatérale).

C'est une comparaison *globale* des distributions, qui peuvent différer par la position, la dispersion, les deux à la fois, ou par toute autre caractéristique.

3.2.2 Conditions d'utilisation

Les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent impérativement être indépendantes pour avoir deux échantillons indépendants. De plus, les variables X_i et Y_j doivent être au moins ordinales, et de préférence continues.

3.2.3 Définition de la statistique, justification intuitive

Soit $\widehat{F}_{m,X}(z)$ la valeur en z de la fonction de répartition empirique du premier échantillon, c'est-à-dire le nombre, divisé par l'effectif m, de valeurs X_i inférieures ou égales à z: c'est une estimation raisonnable de la valeur en z de la fonction de répartition F_X (inconnue) de la première population, car le nombre de valeurs inférieures ou égales à z est une variable binomiale de paramètres m et $F_X(z)$, donc d'espérance $mF_X(z)$. De même, soit $\widehat{F}_{n,Y}(z)$ la valeur en z de la fonction de répartition empirique du second échantillon.

Les statistiques utilisées sont des mesures de l'écart entre ces deux fonctions de répartition empiriques : $D = \sup_z |\widehat{F}_{m,X}(z) - \widehat{F}_{n,Y}(z)|$ pour le test

bilatéral, $D^+ = \sup_z[\widehat{F}_{m,X}(z) - \widehat{F}_{n,Y}(z)]$ pour le premier test unilatéral, et $D^- = \sup_z[\widehat{F}_{n,Y}(z) - \widehat{F}_{m,X}(z)]$ pour le second auront plus tendance à être « trop » grandes sous les alternatives que sous les hypothèses nulles.

3.2.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $D \geqslant d_\alpha$ (table du test bilatéral).
On rejette H_0' au profit de H_1' si $D^+ \geqslant d_\alpha$ (table du test unilatéral).
On rejette H_0'' au profit de H_1'' si $D^- \geqslant d_\alpha$ (table du test unilatéral).

3.2.5 Mise en œuvre

En arrangeant les calculs dans un tableau comme celui de l'exemple (cf. sect. 3.2.7), le calcul de la statistique ne pose aucun problème.

La table donne les valeurs critiques d_α pour m et n entre 5 et 20. Avec de plus grands échantillons, elles sont approchées pour les tests bilatéraux par $1,36\sqrt{(m+n)/mn}$ et $1,63\sqrt{(m+n)/mn}$, et pour les tests unilatéraux par $1,22\sqrt{(m+n)/mn}$ et $1,52\sqrt{(m+n)/mn}$ (aux niveaux 5 % et 1 % resp.).

3.2.6 Compléments

En principe, les variables doivent être continues, mais on peut aussi utiliser le test avec des variables seulement ordinales ; il est alors conservatif, et s'il n'y a que très peu de valeurs distinctes avec beaucoup d'ex æquo (ou si les observations sont groupées en quelques classes) on peut préférer le test χ^2 de la section 3.1 (seul utilisable avec des variables discrètes nominales).

Le test ressemble beaucoup au test de Cramér-von Mises présenté à la section 3.3 suivante, et leurs efficacités semblent comparables : le choix entre les deux relève des préférences individuelles (lorsqu'il y a choix, puisque le test de Cramér-von Mises n'existe qu'en version bilatérale).

Si on l'emploie comme un test de comparaison de positions (ou de dispersions), ce test, comme les autres tests de comparaison *globale* des répartitions, est moins puissant que les tests (de Mann-Whitney-Wilcoxon, de Siegel-Tukey, etc.) spécifiquement adaptés à ces situations particulières (la notion de « stochastiquement plus grand » sous-jacente aux alternatives unilatérales est nettement plus générale qu'un simple décalage de positions).

3.2.7 Exemple

(*Tiré de* Sprent, 1992).

Reprenons l'exemple de la p. 54 : la distribution du nombre de pages est-elle la même pour les livres de statistiques et pour les livres « divers » de la bibliothèque du Pr. Sprent d'où sont tirés les échantillons suivants?

Nombres de pages de 16 livres « divers » :
29 39 60 78 82 112 125 170 192 224 263 275 276 286 369 756
Nombres de pages de 12 livres de statistiques :
126 142 156 228 245 246 370 419 433 454 478 503

Il suffit de calculer les valeurs des fonctions de répartition empiriques $\widehat{F}_X$ et $\widehat{F}_Y$ en chacune des observations (puisqu'elles sont constantes ailleurs). On range les observations par ordre croissant dans les deux premières colonnes du tableau suivant, puis on porte les valeurs de $\widehat{F}_X$ et $\widehat{F}_Y$, qui augmentent par sauts de $1/m$ ou $1/n$ respectivement en chaque X_i ou Y_j observé, et enfin les valeurs absolues de leurs différences.

Livres divers (x_i)	Livres de statistiques (y_j)	$\widehat{F}_X(z)$	$\widehat{F}_Y(z)$	$\lvert\widehat{F}_X(z) - \widehat{F}_Y(z)\rvert$
		(*dans ces trois colonnes, z représente soit x_i soit y_j*)		
29		0,062 5	0	0,062 5
39		0,125	0	0,125
60		0,187 5	0	0,187 5
78		0,25	0	0,25
82		0,312 5	0	0,312 5
112		0,375	0	0,375
125		0,437 5	0	**0,437 5**
	126	0,437 5	0,083 3	0,354 2
	142	0,437 5	0,166 7	0,270 8
	156	0,437 5	0,25	0,187 5
170		0,5	0,25	0,25
192		0,562 5	0,25	0,312 5
224		0,625	0,25	0,375
	228	0,625	0,333 3	0,291 7
	245	0,625	0,416 7	0,208 3
	246	0,625	0,5	0,125
263		0,687 5	0,5	0,187 5
275		0,75	0,5	0,25
276		0,812 5	0,5	0,312 5
286		0,875	0,5	0,375
369		0,937 5	0,5	**0,437 5**
	370	0,937 5	0,583 3	0,354 2
	419	0,937 5	0,666 7	0,270 8
	433	0,937 5	0,75	0,187 5
	454	0,937 5	0,833 3	0,104 2
	478	0,937 5	0,916 7	0,020 8
	503	0,937 5	1	0,062 5
756		1	1	0

On voit facilement que la plus grande différence vaut 0,4375, inférieure à la valeur critique au niveau 5 %, $1/2 = 0,5$, lue dans la table du test bilatéral.

Par conséquent, on ne peut pas rejeter l'hypothèse nulle d'identité des deux fonctions de répartition (alors que le test de comparaison des positions de MANN-WHITNEY-WILCOXON permettait de conclure à un écart statistiquement significatif).

3.2.8 Tables de valeurs critiques d_α du test de Smirnov

Test bilatéral

m \ n :	5	6	7	8	9	10	11	12
5	[1,0|1,0]	4/5	4/5	3/4	7/9	4/5	39/55	43/60
6	1,0	[1,0|5/6]	5/7	17/24	13/18	2/3	43/66	2/3
7	1,0	6/7	[6|6]/7	5/7	2/3	23/35	48/77	53/84
8	7/8	5/6	6/7	[7|6]/8	23/36	3/5	53/88	5/8
9	8/9	5/6	7/9	55/72	[7|6]/9	53/90	59/99	7/12
10	9/10	4/5	53/70	3/4	7/10	[8|7]/10	6/11	11/20
11	9/11	9/11	59/77	8/11	70/99	7/10	[8|7]/11	6/11
12	4/5	5/6	5/7	17/24	25/36	2/3	43/66	[8|7]/12
13	4/5	10/13	5/7	9/13	2/3	42/65	7/11	95/156
14	4/5	16/21	11/14	19/28	2/3	9/14	48/77	13/21
15	4/5	23/30	5/7	27/40	2/3	19/30	34/55	3/5
16	4/5	3/4	11/16	11/16	47/72	5/8	53/88	29/48
17	4/5	73/102	12/17	11/17	11/17	53/85	10/17	7/12
18	7/9	7/9	29/42	47/72	2/3	3/5	59/99	7/12
19	71/95	83/114	13/19	49/76	107/171	113/190	122/209	65/114
20	4/5	11/15	93/140	13/20	37/60	13/20	127/220	17/30

Test bilatéral

m \ n :	13	14	15	16	17	18	19	20
5	9/13	46/70	11/15	54/80	11/17	2/3	61/95	13/20
6	2/3	9/14	19/30	5/8	31/51	2/3	35/57	3/5
7	7/13	9/14	62/105	4/7	4/7	4/7	4/7	79/140
8	31/52	4/7	67/120	5/8	77/136	5/9	41/76	11/20
9	5/9	5/9	5/9	13/24	82/153	5/9	89/171	31/60
10	7/13	37/70	8/15	21/40	89/170	28/45	47/95	11/20
11	75/143	41/77	28/55	89/176	93/187	97/198	102/209	107/220
12	27/52	43/84	31/60	1/2	25/51	1/2	9/19	29/60
13	[9|7]/13	89/182	32/65	101/208	105/221	55/117	6/13	6/13
14	52/91	[9|8]/14	7/15	53/112	111/238	29/63	121/266	9/20
15	23/39	41/70	[9|8]/15	19/40	116/255	41/90	127/285	9/20
16	121/208	9/16	133/240	[5|4]/8	31/68	4/9	7/16	7/16
17	126/221	67/119	142/255	143/272	[10|8]/17	133/306	141/323	73/170
18	131/234	5/9	49/90	77/144	82/153	[10|9]/18	71/171	19/45

suite à la page suivante ...

Test bilatéral

... suite

m	n : 13	14	15	16	17	18	19	20
19	138/247	73/133	152/285	10/19	166/323	88/171	[10|9]/19	8/19
20	11/20	19/35	8/15	21/40	35/68	91/180	187/380	[11|9]/20

Le résultat est significatif si D est supérieur ou égal à la valeur lue dans la table, à [gauche|droite] de la diagonale pour le niveau [1|5] % respectivement.

Test unilatéral

m	n : 5	6	7	8	9	10	11	12
5	[1,0|4/5]	4/5	5/7	27/40	2/3	7/10	7/11	3/5
6	1,0	[1,0|5/6]	2/3	5/8	11/18	3/5	19/33	2/3
7	6/7	5/6	[6|5]/7	17/28	4/7	4/7	4/7	23/42
8	7/8	5/6	3/4	[6|5]/8	5/9	11/20	6/11	13/24
9	4/5	7/9	47/63	3/4	[7|6]/3	5/9	52/99	19/36
10	4/5	11/15	5/7	7/10	61/90	[7|6]/10	57/110	1/2
11	4/5	49/66	5/7	61/88	7/11	69/110	[8|6]/11	16/33
12	4/5	3/4	29/42	2/3	23/36	37/60	7/12	[4|3]/6
13	10/13	9/13	9/13	67/104	73/117	3/5	86/143	23/39
14	51/70	5/7	5/7	9/14	40/63	3/5	45/77	47/84
15	4/5	7/10	23/35	5/8	28/45	3/5	19/33	17/30
16	59/80	11/16	73/112	11/16	29/48	47/80	25/44	9/16
17	63/85	2/3	11/17	5/8	92/153	99/170	104/187	28/51
18	13/18	13/18	83/126	11/18	11/18	26/45	6/11	5/9
19	14/19	77/114	86/133	93/152	11/19	52/95	6/11	121/228
20	3/4	2/3	43/70	5/8	26/45	3/5	59/110	8/15

Test unilatéral

m	n : 13	14	15	16	17	18	19	20
5	8/13	3/5	2/3	3/5	10/17	26/45	56/95	3/5
6	23/39	4/7	17/30	9/16	28/51	11/18	32/57	11/20
7	50/91	4/7	8/15	59/112	61/119	65/126	69/133	18/35
8	27/52	29/56	1/2	9/16	1/2	1/2	37/76	1/2
9	59/117	1/2	23/45	23/48	74/153	1/2	80/171	7/15
10	32/65	17/35	1/2	19/40	79/170	41/90	17/38	1/2
11	67/143	73/154	76/165	5/11	5/11	4/9	92/209	24/55
12	71/156	13/28	7/15	11/24	15/34	4/9	33/76	13/30
13	[8|7]/13	3/7	29/65	7/16	96/221	11/26	8/19	27/65
14	51/91	[8|7]/14	46/105	3/7	50/119	26/63	55/133	57/140
15	107/195	37/70	[9|7]/15	101/240	7/17	37/90	2/5	5/12

suite à la page suivante ...

Test unilatéral

... suite

m \ n :	13	14	15	16	17	18	19	20
16	7/13	15/28	1/2	[9|7]/16	109/272	29/72	15/38	2/5
17	118/221	125/238	131/255	139/272	[9|8]/17	59/153	126/323	13/34
18	41/78	65/126	23/45	71/144	25/51	[5|4]/9	7/18	17/45
19	10/19	135/266	142/285	151/304	158/323	80/171	[10|8]/19	36/95
20	27/52	71/140	1/2	39/80	163/340	17/36	9/20	[5|4]/10

Le résultat est significatif si D^+ (ou D^- qui suit la même distribution) est supérieur ou égal à la valeur lue dans la table, à [gauche|droite] de la diagonale (ou en [dessous|dessus]) pour le niveau [1|5] % respectivement.

3.2.9 Mise en œuvre informatique

Il n'y a pas de fonction SPLUS pour le test de Cramér-von Mises.

SAS L'option EDF de la procédure NPAR1WAY permet d'obtenir les statistiques non paramétriques sur les tests de distribution.

```
data lost;
     input livre $ npages @@;
    cards;
R 29 R 39 R 60 R 78 R 82 R 112 R 125 R 170 R 192
R 224 R 263 R 275 R 276 R 286 R 369 R 756
S 126 S 142 S 156 S 228 S 245 S 246 S 370
S 419 S 433 S 454 S 478 S 503
;
proc npar1way EDF;
     class livre;
     var npages;
run;
```

Avec les commandes ci-dessus, SAS réalise aussi le test de Cramér-von-Mises. Ci-dessous, on ne donne que la sortie pour le test de Smirnov. La sortie pour le test de Cramér-von-Mises est donnée dans la section suivante.

```
          N P A R 1 W A Y  P R O C E D U R E

     Kolmogorov-Smirnov Test for Variable NPAGES
          Classified by Variable LIVRE

                                          Deviation
                          EDF             from Mean
   LIVRE       N       at maximum        at maximum
   R          16      0.437500000       0.750000000
   S          12      0.000000000       -.866025404
   -----      --      -----------
```

```
          28      0.250000000

Maximum Deviation occurred at Observation    7
Value of NPAGES   at maximum  125.000000

     Kolmogorov-Smirnov 2-Sample Test (Asymptotic)
          KS = 0.216506      D = 0.437500
      KSa =  1.14564      Prob > KSa =   0.1448
```

3.3 Test de Cramér-von Mises

3.3.1 Définition du problème: hypothèses nulle et alternative

Tester l'identité des fonctions de répartition F_X et F_Y des populations d'où sont tirés les échantillons $\{X_i,\ i = 1, \ldots, m\}$ et $\{Y_j,\ j = 1, \ldots, n\}$, contre la seule alternative bilatérale. On teste donc :

H_0 : $F_X(z) = F_Y(z)$ pour tout réel Z, contre

H_1 : $F_X(z) \neq F_Y(z)$ pour au moins un réel z.

C'est une comparaison *globale* des distributions, qui peuvent différer aussi bien par la position que par la dispersion, les deux à la fois, ou par toute autre caractéristique.

3.3.2 Conditions d'utilisation

Les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent impérativement être indépendantes pour avoir deux échantillons indépendants.

De plus, les variables X_i et Y_j doivent être au moins ordinales, et de préférence continues.

3.3.3 Définition de la statistique, justification intuitive

Soit $\widehat{F}_{m,X}(z)$ la valeur en z de la fonction de répartition empirique du premier échantillon, c'est-à-dire le nombre, divisé par l'effectif m, de valeurs X_i inférieures ou égales à z : c'est une estimation raisonnable de la valeur en z de la fonction de répartition F_X (inconnue) de la première population, car le nombre de valeurs inférieures ou égales à z est une variable binomiale d'effectif m et de probabilité $F_X(z)$, dont l'espérance vaut donc $mF_X(z)$.

De même, soit $\widehat{F}_{n,Y}(z)$ la valeur en z de la fonction de répartition empirique du second échantillon. La statistique utilisée mesure la différence entre les deux fonctions de répartition inconnues par la somme des carrés des écarts entre les valeurs de leurs versions empiriques aux points observés : $D^2 = \sum_{\{z=X_i \text{ ou } Y_j\}} [\widehat{F}_{m,X}(z) - \widehat{F}_{n,Y}(z)]^2$.

3.3.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 au niveau 5 % ou 1 % si $T = mnD^2/(m+n)^2$ dépasse 0,461 ou 0,743 respectivement.

3.3.5 Mise en œuvre

En arrangeant les calculs dans un tableau comme celui de l'exemple (cf. sect. 3.3.7), analogue à l'exemple du test de SMIRNOV (sect. 3.2.7), le calcul de la statistique ne pose aucun problème. Il est néanmoins plus long que celui de la statistique de SMIRNOV.

Les valeurs critiques données à la section précédente sont une approximation asymptotique. En pratique, cette approximation est suffisamment bonne, même avec de très petits échantillons, pour éviter le recours à des tables.

3.3.6 Compléments

Le test ressemble beaucoup au test de SMIRNOV présenté à la section 3.2 précédente, et leurs efficacités semblent comparables : lorsqu'il y a choix entre les deux (remarquer que, contrairement au test de SMIRNOV, le test de CRAMÉR-VON MISES n'existe qu'en version bilatérale), ce choix relève des préférences individuelles.

3.3.7 Exemple

(*Tiré de* SPRENT, 1992).

Reprenons l'exemple de la p. 54 : la distribution du nombre de pages est-elle identique pour les deux « populations » de livres (livres de statistiques et livres « divers ») de la bibliothèque du Pr. SPRENT d'où sont tirés les deux échantillons suivants ?

Nombres de pages de 16 livres « divers » :
29 39 60 78 82 112 125 170 192 224 263 275 276 286 369 756
Nombres de pages de 12 livres de statistiques :
126 142 156 228 245 246 370 419 433 454 478 503

Il suffit de calculer les valeurs des fonctions de répartition empiriques $\widehat{F}_X$ et $\widehat{F}_Y$ en chacune des observations (puisqu'elles sont constantes ailleurs).

On range les observations par ordre croissant dans les deux premières colonnes du tableau suivant, puis on porte les valeurs de $\widehat{F}_X$ et $\widehat{F}_Y$, qui augmentent par sauts de $1/m$ ou $1/n$ respectivement en chaque X_i ou Y_j observé.

Enfin, on porte dans la dernière colonne les carrés de leurs différences et on calcule la somme de ces carrés.

Livres divers (X_i)	Livres de statistiques (Y_j)	$\widehat{F}_X(z)$	$\widehat{F}_Y(z)$	$\left[\widehat{F}_X(z) - \widehat{F}_Y(z)\right]^2$
		(*dans ces trois colonnes, z représente soit x_i soit y_j*)		
29		0,0625	0	0,0039
39		0,125	0	0,0156
60		0,1875	0	0,0352
78		0,25	0	0,0625
82		0,3125	0	0,0977
112		0,375	0	0,1406
125		0,4375	0	0,1914
	126	0,4375	0,0833	0,1254
	142	0,4375	0,1667	0,0734
	156	0,4375	0,25	0,0352
170		0,5	0,25	0,0625
192		0,5625	0,25	0,0977
224		0,625	0,25	0,1406
	228	0,625	0,3333	0,0851
	245	0,625	0,4167	0,0434
	246	0,625	0,5	0,0156
263		0,6875	0,5	0,0352
275		0,75	0,5	0,0625
276		0,8125	0,5	0,0977
286		0,875	0,5	0,1406
369		0,9375	0,5	0,1914
	370	0,9375	0,5833	0,1254
	419	0,9375	0,6667	0,0734
	433	0,9375	0,75	0,0352
	454	0,9375	0,8333	0,0109
	478	0,9375	0,9167	0,0004
	503	0,9375	1	0,0039
756		1	1	0

On trouve ainsi $D^2 \approx 2,002$ et donc $T = 12 \times 16 \times 2,002/(28)^2 = 0,490$. Comme cette statistique dépasse la valeur critique 0,461 du niveau 5 %, on rejette à ce niveau l'hypothèse de distributions identiques pour les livres de statistique et les livres divers.

3.3.8 Table de valeurs critiques

Comme signalé plus haut, la bonne qualité de l'approximation asymptotique permet de s'en dispenser (il semble que l'on n'ait calculé de valeurs exactes que pour des tailles d'échantillons très faibles, pour $m + n \leqslant 17$).

3.3.9 Mise en œuvre informatique

SAS L'option EDF de la procédure NPAR1WAY permet d'obtenir les statistiques non paramétriques sur les tests de distribution.

```
data lost;
     input livre $ npages @@;
    cards;
R 29 R 39 R 60 R 78 R 82 R 112 R 125 R 170 R 192
R 224 R 263 R 275 R 276 R 286 R 369 R 756
S 126 S 142 S 156 S 228 S 245 S 246 S 370
S 419 S 433 S 454 S 478 S 503
;
proc npar1way EDF;
     class livre;
     var npages;
run;
```

Avec les commandes ci-dessus, SAS réalise aussi le test de Smirnov. Ci-dessous, on ne donne que la sortie pour le test de Cramér-von-Mises. La sortie pour le test de Smirnov est donnée dans la section précédente.

```
                    N P A R 1 W A Y  P R O C E D U R E

            Cramer-von Mises Test for Variable NPAGES
                 Classified by Variable LIVRE

                                              Summed
                                            Deviation
LIVRE                    N                  from Mean

R                       16                0.210140306
S                       12                0.280187075


              Cramer-von Mises Statistic (Asymptotic)
                 CM = 0.017512      CMa = 0.490327
```

On compare cette statistique asymptotique CMa au seuil critique 0.461 (5%) ou 0.743 (1%).

3.4 Test des runs de Wald-Wolfowitz

3.4.1 Définition du problème: hypothèses nulle et alternative

Tester l'identité des fonctions de répartition F_X et F_Y des populations d'où sont tirés les échantillons $\{X_i,\ i = 1, \ldots, m\}$ et $\{Y_j,\ j = 1, \ldots, n\}$, contre la seule alternative bilatérale. On teste donc :

H_0 : $F_X(z) = F_Y(z)$ pour tout réel Z, contre

H_1 : $F_X(z) \neq F_Y(z)$ pour au moins un réel z.

C'est une comparaison *globale* des distributions, qui peuvent différer aussi bien par la position que par la dispersion, les deux à la fois, ou par toute autre caractéristique.

3.4.2 Conditions d'utilisation

Les variables $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ doivent impérativement être indépendantes pour avoir deux échantillons indépendants.

De plus, les variables X_i et Y_j doivent être au moins ordinales, et de préférence continues.

3.4.3 Définition de la statistique, justification intuitive

On range ensemble les deux échantillons par ordre croissant, et on remplace chaque valeur par le numéro de son échantillon d'origine. On obtient ainsi une suite de $m + n$ symboles 1 ou 2.

On définit un *run* comme une séquence d'observations de l'un des deux échantillons (une séquence de symboles 1 ou de 2) immédiatement suivie soit d'une séquence d'observations de l'autre échantillon, soit de rien (fin de la suite des $m + n$ symboles).

La statistique utilisée est le nombre total R de runs dans la suite construite à partir des deux échantillons. Sous H_0, les X_i et les Y_j sont tous issus d'une même population et on ne doit pas s'attendre à observer un nombre de runs trop faible, contrairement à certaines alternatives (cas limites : seulement 2 runs si la différence des positions est très forte par rapport aux dispersions — par exemple *111111-222222* ; ou 3 runs si la position est commune mais si les dispersions sont très différentes — par exemple *111-222222-111*).

3.4.4 Régions de rejet et de non-rejet

On rejette H_0 au profit de H_1 si $R \leqslant R^i_\alpha$, la valeur critique inférieure du test unilatéral de runs sur la nature d'un échantillon vu en section 2.1.

3.4.5 Mise en œuvre

Le calcul de la statistique ne pose pas de problème sauf s'il y a des ex æquo : dans le cas où certains X_i sont égaux à certains Y_j, il faut établir la suite de symboles de façon à rendre R le plus grand possible (résolution *conservative* des ex æquo, c'est-à-dire « la plus favorable à H_0 »). Par exemple, avec les données suivantes :

X : 1 *4* 5 8 9 $(m = 5)$
Y : 2 3 *4* 6 7 10 $(n = 6)$,

la suite réordonnée des X_i et Y_j sera :

1 2 3 *4* *4* 5 6 7 8 9 10

que l'on peut transformer en suite de 1 et 2 des deux manières suivantes :

1 2 2 2 1 1 2 2 1 1 2 où $R = 6$, et :
1 2 2 1 2 1 2 2 1 1 2 où $R = 8$,
et l'on retiendra donc cette dernière valeur.

3.4.6 Compléments

Ce test est en fait souvent moins puissant que ses concurrents de SMIRNOV (cf. sect. 3.2) ou de CRAMÉR-VON MISES (cf. sect. 3.3) et n'a donc que peu d'intérêt pratique.

3.4.7 Exemple

Un psychologue note le temps (en secondes) nécessaire à la réalisation d'un exercice manuel, pour 7 enfants considérés comme non-handicapés et 8 enfants considérés comme handicapés. Les résultats sont les suivants :

« Non-handicapés » : 204 218 197 183 227 233 191 $(m = 7)$
« Handicapés » : 243 228 261 202 343 242 220 239 $(n = 8)$

Le test de WALD-WOLFOWITZ permet-il de conclure que les deux échantillons proviennent de populations différentes ?

En ordonnant les deux échantillons ensemble, on obtient la suite :
1 1 1 2 1 1 2 1 2 1 2 2 2 2 2
et donc $R = 8$ runs. On voit dans la table que la valeur critique au niveau 5 % est 4. Par conséquent le test ne permet pas de conclure à une différence entre les deux populations.

3.4.8 Table

Se reporter à la table de la section 2.1.8 (p. 23), et prendre à gauche de la diagonale la plus petite des deux valeurs de la case adéquate.