Cours sur le logiciel 

M2 - Recherche
Mouvement, Performance, Santé

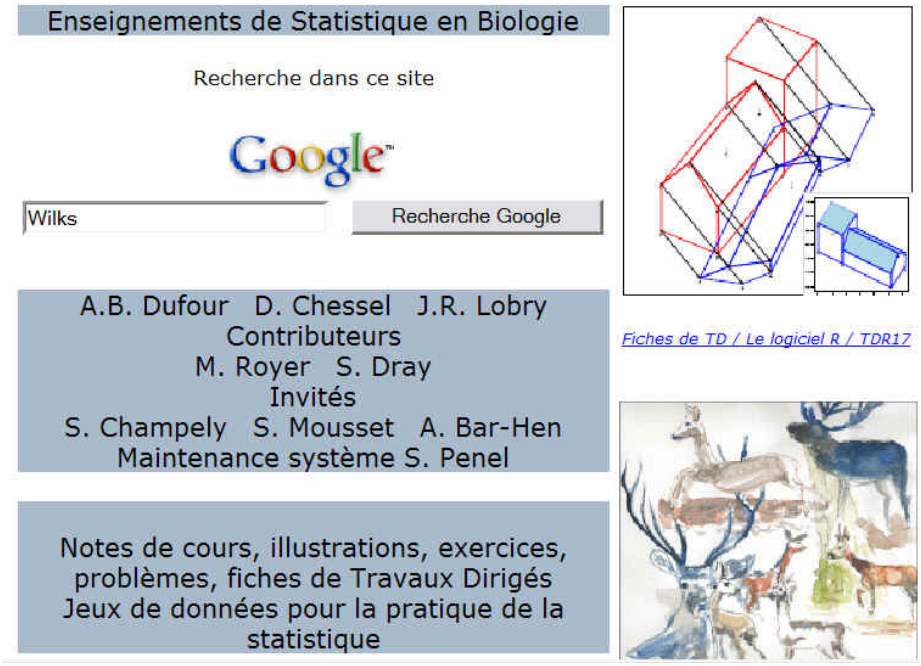
Anne B. Dufour

11 septembre 2006

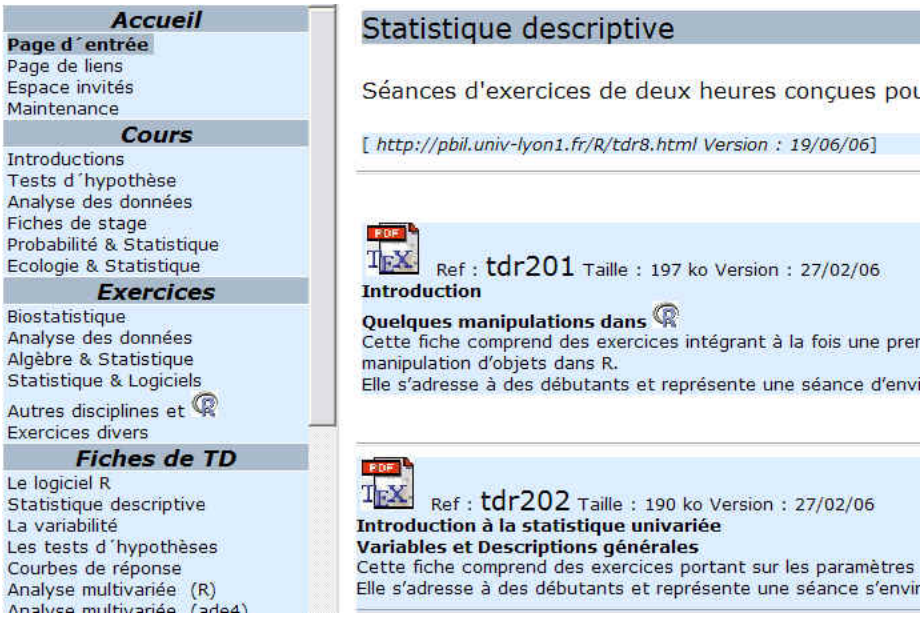
Table des matières

1 Introduction

Enseigner la statistique est un exercice difficile. Cela relève autant de la théorie mathématique que de l'application. Ceux qui ont besoin de la statistique n'en n'ont pas la maîtrise, ceux qui la conçoivent n'en n'ont pas l'usage. Le document apparaîtra pour certains comme une succession d'outils. C'est le cas mais pas notre souhait. Nous l'avons conçu comme une base nécessaire à tous. C'est pourquoi nous donnons de nombreuses références pour que chacun puisse aller puiser d'autres informations, parfois plus simple, parfois plus complexe. Le site pédagogique `www.pbil.univ-lyon1.fr/R/enseignement.html` sera souvent cité et nous vous engageons vivement à vous y rendre.



Le site comprend un menu déroulant et un google interne qui vous permettront de rechercher d'autres tests, d'autres méthodes. C'est la question scientifique, quelle que soit sa nature, qui prime et non la boîte à outils qui vous est mis à disposition.



2 Ajustement à une loi théorique

2.1 Test d'ajustement du Chi-Deux

2.1.1 Enoncé pratique du test

Soit X une variable aléatoire, prenant p modalités (dans le cas où elle est discrète) ou p classes d'intervalles (dans le cas où elle est continue), étudiée sur un échantillon de taille n. On s'intéresse à l'ajustement de X à une loi théorique T.

Hypothèses :
- H_0 : X suit la loi théorique T
- H_1 : X ne suit pas la loi théorique T.

Valeur de la Statistique du test :

$$\chi^2 = \sum \frac{(n_i - np_i)^2}{np_i}$$

où les n_i représentent les effectifs observés et les np_i les effectifs théoriques.

Sous H_0, χ^2 suit une loi du Chi-Deux à ν=$(p - c)$ degrés de liberté c'est-à-dire le nombre de composantes moins le nombre de relations qui les lient.

Exemples
- Ajustement à une loi uniforme : c = 1 (valeur de n) donc $\nu = p - 1$
- Ajustement à une loi binomiale : c = 2 (valeurs de n et de p) donc $\nu = p - 2$

Décision statistique au seuil α :
Soit la valeur $\chi^2_{1-\alpha}$, lue dans la table du Chi-Deux.
- si $\chi^2 \in \left[\chi^2_{1-\alpha}\,;\,+\infty\right[$, on rejette H_0.
- si $\chi^2 \notin \left[\chi^2_{1-\alpha}\,;\,+\infty\right[$, on accepte H_0.

2.1.2 Exemple

On demande à 50 dégustateurs de vin rosé de choisir parmi 5 vins, celui qu'ils préfèrent. En fait, c'est le même vin mais servi dans des verres de couleurs différentes (du plus foncé (1) au plus clair (5)). La couleur du verre influence-t-elle le choix du vin ?

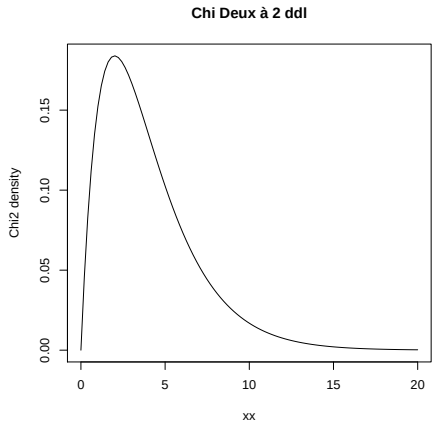
Hypothèses :
- H_0 : La couleur du verre n'influence pas le choix du vin.
- H_1 : La couleur du verre influence le choix du vin.

Ces hypothèses peuvent se traduire d'une manière plus statistique.
- H_0 : La distribution suit une loi uniforme.
- H_1 : La distribution ne suit pas une loi uniforme.

Tableau des données et des calculs :

vin i	1	2	3	4	5
observé n_i	6	12	9	10	13
théorique np_i	10	10	10	10	10

Statistique du test : $\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$ soit $\chi^2 = 3$
Sous H_0, la statistique du test suit une loi du Chi-Deux à 4 ddl représentée ci-dessous.



Décision statistique au risque α=0.05 :

```
qchisq(0.95, 4)
```

```
[1] 9.487729
```

$\chi^2 \notin \left[\chi^2_{1-\alpha}\,;\,+\infty\right[$, donc on ne peut pas dire que la couleur du verre influence le choix du vin.

L'utilisation du logiciel R facilitant le calcul des probabilités des grandes lois théoriques permet de tenir un autre raisonnement quant à la décision prise au cours d'un test. Nous pouvons en effet, calculer la probabilité exacte associée à la valeur calculée de la statistique du test (dans notre cas 3).

```
1 - pchisq(3, 4)
```

```
[1] 0.5578254
```

Cette probabilité est appelée `p-value`. Nous pouvons dire alors que si $p < \alpha$, l'hypothèse H_0 est rejetée. Ce résultat se retrouve directement.

Test sous R

```
chisq.test(c(6, 12, 9, 10, 13), p = rep(1/5, 5))
```

```
        Chi-squared test for given probabilities

data:  c(6, 12, 9, 10, 13)
X-squared = 3, df = 4, p-value = 0.5578
```

Conclusion : La couleur du verre n'influence pas le choix du vin.

2.2 Test d'ajustement de Kolmogorov-Smirnov

2.2.1 Enoncé pratique du test

Soit $F(x)$ la fonction de répartition d'une variable aléatoire continue X.

Soit $F_n(x)$, la fonction de répartition empirique correspondant à un échantillon de taille n.

Le test de Kolmogorov-Smirnov est basé sur la comparaison de la fonction de la fonction de répartition des données de l'échantillon avec la fonction de répartition $F(x)$ de la population.

Hypothèses :

- H_0 : X suit la loi théorique T.
- H_1 : X ne suit pas la loi théorique T.

Statistique du test : $D_n = Sup_X\,|F_n(x) - F(x)|$

L'écart entre observé et théorique est mesuré en chaque point.

Sous H_0, D_n suit une loi de Kolmogorov-Smirnov pour un effectif total n.

Décision statistique au risque α :

Soit la valeur D_t, lue dans la table de Kolmogorov-Smirnov.
- si $D_n \in [D_t\,;\ +\infty[$, on rejette H_0.
- si $D_n \notin [D_t\,;\ +\infty[$, on accepte H_0.

Remarques :

- Dans le cas d'un ajustement à une loi normale, on utilise les paramètres de la population (moyenne μ et variance σ^2) lorsqu'ils sont connus ; leurs estimations dans le cas contraire. La table de référence n'est plus alors celle de Kolmogorov-Smirnov mais celle de Lilliefors.
- Dans le cas d'un ajustement à une loi normale, on préfèrera le test de Shapiro-Wilks (cf détail dans la fiche `tdr31. Comparaison de moyennes`).

2.2.2 Exemple

Les données sont extraites de l'enquête longitudinale du Pr G. Beunen, de l'université Catholique de Leuven (Belgique). 28 hommes ont été choisis au hasard et nous avons relevé pour chacun d'eux les résultats à trois tests d'aptitudes physiques à 18 ans puis à 30 ans.

- VTJ : Vertical Jump ou Détente Verticale (en centimètres)
- ARM : Arm Pull ou Mesure de la force statique du bras (en kilogrammes)
- SHR : Shuttle Run ou Course navette (10 x 5 mètres, en dixièmes de seconde)
- PO : Pulse ou Nombre de pulsations par minutes

Prenons par exemple la variable *Détente verticale* à l'âge de 18 ans.

```
aptmot = read.table("aptmot.txt", h = T)
vtj = aptmot$VTJ18
summary(vtj)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   45.00   48.00   49.89   54.25   71.00
```

```
var(vtj)
```

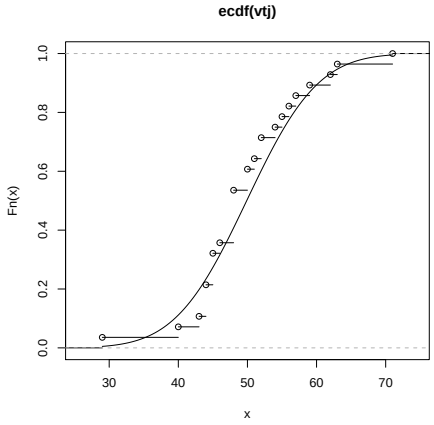
```
[1] 65.87698
```

```
sd(vtj)
```

```
[1] 8.116464
```

Le test étant construit sur la fonction de répartition des observations cumulées, la représentation graphique la plus appropriée est la suivante (ecdf : empirical cumulative distribution function).

```
plot(ecdf(vtj))
xx = seq(29, 71, le = 100)
lines(xx, pnorm(xx, mean = mean(vtj), sd = sd(vtj)))
```



Test sous R

```
ks.test(vtj, "pnorm", mean(vtj), sd(vtj))
```

```
        One-sample Kolmogorov-Smirnov test

data:  vtj
D = 0.1279, p-value = 0.7493
alternative hypothesis: two.sided
```

Test sous R

```
shapiro.test(vtj)
```

```
        Shapiro-Wilk normality test

data:  vtj
W = 0.9542, p-value = 0.2526
```

2.3 Exercice

Reprenons le fichier des données de l'enquête Belge.

a) Représenter le nuage de points de la détente verticale à 18 ans (en abscisse) et de la détente verticale à 30 ans (en ordonnées)

b) Tracer à l'aide de la commande `abline` la première bissectrice c'est-à-dire la droite $y = x$.

c) Commenter le résultat.

d) Réitérer cette démarche avec la variable "Récupération : PO".

e) Représenter pour le pouls à 18 ans (resp. à 30 ans) les deux fonctions de répartition empirique et théorique. On choisira comme distribution théorique celle de la loi normale.

f) Commenter.

3 Deux tests de comparaison de variances

3.1 Enoncé pratique du test

Soit X une variable continue mesurée dans p échantillons, indépendants entre eux, d'effectifs $n_1, n_2, \ldots, n_p$. Le nombre total d'observations est $n = \sum_{k=1}^{p} n_k$. Ces échantillons sont issus de p populations. La variable est normalement distribuée.

On note $\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2$ les variances inconnues dans les p populations.

On note $\widehat{\sigma_1^2}, \widehat{\sigma_2^2}, \cdots, \widehat{\sigma_p^2}$, les variances estimées et $s_1^2, s_2^2, \cdots, s_p^2$ les variances descriptives.

$$\widehat{\sigma_k^2} = \frac{1}{(n_k-1)} \sum_{i=1}^{n_k} (x_{ik} - \overline{x_k})^2 \text{ et } s_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{ik} - \overline{x_k})^2$$

Hypothèse Nulle :
" La variable X a la même variance dans les p populations " :

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$$

3.1.1 Test de Bartlett

La valeur de la statistique du test est :

$$B = (n-p)Log\widehat{\sigma^2} - \sum_{k=1}^{p} (n_k - 1)Log\widehat{\sigma_k^2} \text{ où } \widehat{\sigma^2} = \frac{1}{n-p} \sum_{k=1}^{p} (n_k - 1)\widehat{\sigma_k^2}$$

Sous l'hypothèse H_0, la statistique de Bartlett suit une loi du Chi-Deux à (p-1) degrés de liberté.

On peut améliorer la conformité de la statistique à la loi du Chi-Deux en réalisant la transformation suivante :

$$c = 1 + \frac{1}{3(p-1)} \left[\sum_{k=1}^{p} \left(\frac{1}{n_k - 1} \right) - \frac{1}{n-p} \right]$$

La valeur de la statistique de Bartlett devient alors : $B_c = \frac{B}{c}$.

3.1.2 Test de Cochran

Les échantillons sont tous de taille identique $n = n_1 = n_2 = ... = n_p$
La valeur de la statistique du test est :

$$C = \frac{\max\left(\widehat{\sigma_1^2}, \widehat{\sigma_2^2}, \cdots, \widehat{\sigma_p^2}\right)}{\sum_{k=1}^{p} \widehat{\sigma_k^2}}$$

La décision se fait par rapport à une valeur critique $C_\alpha(n-1,p)$ que l'on trouve dans une table spécifique. Et on rejette l'hypothèse lorsque $C > C_\alpha(n-1,p)$. Ce test n'est pas programmé sous R.

[Equivalence de la table : $k = p$ et $\nu_x = n - 1$]

3.1.3 Remarque

Le test de Cochran, comme le test de Bartlett est très sensible à la non normalité de la variable. C'est pourquoi, on peut préférer le test de Levene (à condition d'être dans le cas de grands échantillons) ou le test " log-anova ".

3.2 Exemple

On connaît le salaire pour 18 sportifs américains choisis parmi les sportifs les mieux payés de la planète (extrait du magazine américain Forbes).

basket	boxe	baseball
31,3	54,3	13,2
17	38	13
13,2	27	11
13	14,7	10,6
11	12	9,5
10,4	9,7	9,3

```
salaires <- c(31.3, 17, 13.2, 13, 11, 10.4, 54.3, 38,
    27, 14.7, 12, 9.7, 13.2, 13, 11, 10.6, 9.5, 9.3)
length(salaires)
```

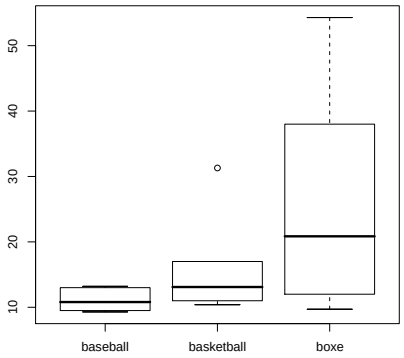
```
[1] 18
```

```
sport <- rep(c("basketball", "boxe", "baseball"), rep(6,
    3))
sport <- factor(sport)
boxplot(salaires ~ sport)
tapply(salaires, sport, mean)
```

```
 baseball basketball       boxe
 11.10000   15.98333   25.95000
```

```
tapply(salaires, sport, var)
```

```
 baseball basketball       boxe
  2.81600   61.65767  307.05100
```



a) Test de Bartlett

```
bartlett.test(salaires, sport)
```

```
        Bartlett test of homogeneity of variances

data:  salaires and sport
Bartlett's K-squared = 16.407, df = 2, p-value = 0.0002737
```

b) Test de Cochran

```
vargroup <- tapply(salaires, sport, var)
max(vargroup)/sum(vargroup)
```

```
[1] 0.826462
```

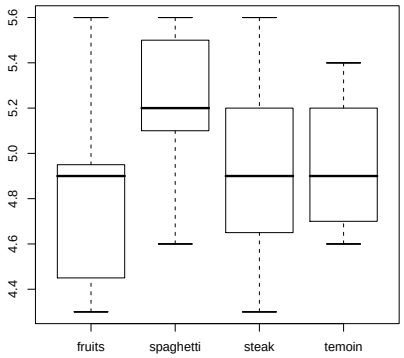
Si nous utilisons la formule qui permet d'améliorer la convergence vers la loi du Chi-Deux, nous obtenons C=0.8264619.

Lecture de la table du test de Cochran : $C_{0.05}(5,3) = 0.7071$.

3.3 Exercice

" Pour déterminer si le régime affecte le débit cardiaque (en l/min) chez des personnes vivant dans une petite ville, nous avons sélectionné au hasard quatre groupes de sept personnes chacun. Les sujets du groupe témoin continuait à manger normalement ; ceux du second groupe ne mangeaient que des spaghetti ; ceux du troisième groupe ne mangeaient que des steaks ; ceux du quatrième groupe ne mangeaient que des noix et des fruits. " (in Introduction aux biostatistiques, S.A.Glantz, McGraw-Hill, 1998).

témoin	spaghetti	steak	fruits/noix
4,6	4,6	4,3	4,3
4,7	5	4,4	4,4
4,7	5,2	4,9	4,5
4,9	5,2	4,9	4,9
5,1	5,5	5,1	4,9
5,3	5,5	5,3	5
5,4	5,6	5,6	5,6



4 Analyse de la variance à un facteur

4.1 Modèle d'analyse de la variance

On cherche à expliquer une variable quantitative X par une variable explicative de type qualitatif à p modalités. Le modèle étudié est le suivant : $X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

- ⋆ X_{ij} est la valeur de la variable quantitative pour un individu i de la modalité j.
- ⋆ μ est la moyenne totale.
- ⋆ α_j est l'écart entre les valeurs de la modalité j et la moyenne générale.
- ⋆ ε_{ij} représente les erreurs indépendantes, d'espérance nulle et de variance constante.

4.2 Principe

L'hypothèse à tester concerne l'égalité des moyennes dans les p populations caractérisées par la variable qualitative :

H_0 : le caractère a la même moyenne dans les p populations.

$$\mu_1 = \mu_2 = ... = \mu_p$$

Le critère à calculer est : $F = \dfrac{(n-p)\sum_{k=1}^{p} n_k(\bar{x}_k - \bar{x})^2}{(p-1)\sum_{k=1}^{p} n_k s_k^2}$

où $\bar{x}$ est la moyenne générale, $\bar{x}_k$et s_k^2 respectivement la moyenne et la variance des valeurs constituant le groupe k.

Les résultats se présentent généralement sous la forme d'un tableau appelé tableau d'analyse de la variance.

Source de variation	Variation	Ddl	Carré moyen	F
Inter groupes	$SCE_B = \sum_{k=1}^{p} n_k(\bar{x}_k - \bar{x})^2$	p-1	$CM_B = \frac{SCE_B}{p-1}$	$F = \frac{CM_B}{CM_W}$
Intra groupes	$SCE_W = \sum_{k=1}^{p} n_k s_k^2$	$n-p$	$CM_W = \frac{SCE_W}{n-p}$	
Totale	$SCE_T = \sum_{i=1}^{n} (x_i - \bar{x})^2$	n-1		

Si le caractère observé est distribué selon une loi de Gauss dans chacune des p populations (contrainte de normalité), et si les variances dans les p populations sont égales (contrainte d'homoscédasticité), alors, sous l'hypothèse H_0, le critère F est la valeur observée d'une variable aléatoire distribuée selon une loi de Fisher - Snedecor à $p-1$ et $n-p$ degrés de liberté.

La contrainte de normalité peut être vérifiée au préalable par un test d'ajustement à une loi de Gauss (test de Kolmogorov-Smirnov, de Lilliefors, de Shapiro-Wilks). La contrainte d'homoscédasticité peut être vérifiée au préalable par une test de Bartlett) ou par un test de Cochran.

4.3 Exemple

33 sportifs en fauteuil roulant évoluant à des niveaux de compétition différents : international, national, régional et récréatif ont accepté de passer un test psychologique. Une des mesures réalisées porte sur l'anxiété des sportifs au moment de la compétition.

```
psycho <- read.table("psycho.txt", h = T)
anxiete <- psycho$anxiete
niveau <- factor(psycho$niveau)
summary(anxiete)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.80   33.40   39.10   38.72   44.50   51.20
```

```
var(anxiete)
```

```
[1] 54.86189
```

```
summary(niveau)
```

```
international      national     recreatif      regional
            8             8             9             8
```

```
tapply(anxiete, niveau, mean)
```

```
international      national     recreatif      regional
     32.95000      38.03750      41.52222      42.03750
```

```
tapply(anxiete, niveau, var)
```

```
international      national     recreatif      regional
     46.36857      45.21696      41.51444      50.51125
```

```
 boxplot(anxiete ~ niveau)
 bartlett.test(anxiete, niveau)
```

```
        Bartlett test of homogeneity of variances

data:  anxiete and niveau
Bartlett's K-squared = 0.0692, df = 3, p-value = 0.9953
```

```
 shapiro.test(anxiete[niveau == "international"])
```

```
        Shapiro-Wilk normality test

data:  anxiete[niveau == "international"]
W = 0.9212, p-value = 0.4399
```

```
 shapiro.test(anxiete[niveau == "national"])
```

```
        Shapiro-Wilk normality test

data:  anxiete[niveau == "national"]
W = 0.968, p-value = 0.8822
```

```
 shapiro.test(anxiete[niveau == "recreatif"])
```

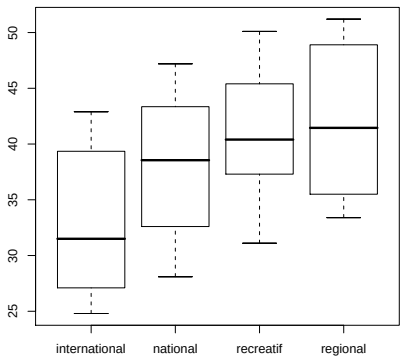
```
        Shapiro-Wilk normality test

data:  anxiete[niveau == "recreatif"]
W = 0.9591, p-value = 0.7886
```

```
 shapiro.test(anxiete[niveau == "regional"])
```

```
        Shapiro-Wilk normality test

data:  anxiete[niveau == "regional"]
W = 0.9051, p-value = 0.3205
```



L'analyse de la variance à un facteur contrôlé (ANOVA1), permet de tester l'hypothèse H_0 :

"En moyenne, l'anxiété est la même quel que soit le niveau de compétition du sportif handicapé."

Test sous R

```
anova(lm(anxiete ~ niveau))
```

```
Analysis of Variance Table

Response: anxiete
          Df  Sum Sq Mean Sq F value  Pr(>F)
niveau     3  428.79  142.93   3.124 0.04101 *
Residuals 29 1326.79   45.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4 Exercice

Reprenons l'exemple du débit cardiaque en fonction du régime alimentaire suivi (3.3). Peut-on mettre en évidence, en moyenne, une différence du débit cardiaque en fonction du régime alimentaire ?

5 Test de Kruskal-Wallis

5.1 Principe

L'hypothèse à tester concerne un caractère mesurable (continu, discret, ou ordinal), observé dans p échantillons, indépendants entre eux, d'effectifs n_1, n_2, ..., n_p. (tous les n_p sont strictement supérieurs à 5 ; le nombre total d'observations est $n = \sum_{k=1}^{p} n_k$.

H_0 : "le caractère est distribué de manière identique dans les p populations"

A chaque valeur observée correspond son rang, dans la liste des n valeurs classées dans l'ordre croissant. Si q valeurs sont égales (ex-æquo), on attribue à chacune le rang moyen (moyenne arithmétique du rang qu'elles occupent dans la liste des rangs de 1 à n).

Le critère à calculer est :

$$H = \frac{12}{n(n+1)} \sum_{k=1}^{p} \frac{R_k^2}{n_k} - 3(n+1)$$

où R_k note la somme des rangs des valeurs constituant le groupe k. S'il existe m ensembles d'ex-æquo, la quantité H ci-dessus doit être corrigée, en la divisant par :

$$c = 1 - \frac{\sum_{e=1}^{m} (q_e - 1) q_e (q_e + 1)}{n^3 - n}$$

où q_e est le nombre de valeurs (égales entre elles) constituant l'ensemble e d'ex-æquo.
On a alors : $Hc = \frac{H}{c}$

Sous l'hypothèse H_0, le critère H (éventuellement corrigé Hc) est la valeur observée d'une variable aléatoire distribuée selon une loi de χ^2 à p-1 degrés de liberté.

5.2 Exemple

Reprenons l'exemple du salaire des 18 sportifs américains choisis parmi les sportifs les mieux payés de la planète (3.2). Les boîtes à moustaches ont montré la forte variabilité des salaires et les tests de variances ont montré que l'hypothèse d'hétéroscédasticité était vérifiée. On réalise donc un test de Kruskal-Wallis.

Test sous R

```
kruskal.test(salaires, sport)
```

```
        Kruskal-Wallis rank sum test

data:  salaires and sport
Kruskal-Wallis chi-squared = 4.5257, df = 2, p-value = 0.1041
```

5.3 Exercices

5.3.1 Anxiété chez les sportifs

Reprenons l'exemple de la mesure de l'anxiété chez 33 sportifs en fauteuil roulant appartenant à quatre niveaux de pratique différents (4.3).

a) Donner l'hypothèse nulle associée à cette situation ?

b) Réaliser un test de Kruskal-Wallis.

c) Si vous aviez le choix entre plusieurs tests, que feriez-vous ? pourquoi ?

5.3.2 Débit cardiaque et régime alimentaire

Mêmes questions à propos du débit cardiaque en fonction du régime alimentaire.

6 Test du Chi-Deux de Contingence

6.1 La table de contingence observée

Soient A et B, deux variables qualitatives ayant respectivement p et q modalités. Soit n, le nombre d'individus sur lesquels A et B ont été observées.

La *table de contingence* observée est la construction d'un tableau croisé où les colonnes correspondent aux q modalités de la variable B et les lignes, aux p modalités de la variable A. On note n_{ij} le nombre d'individus possédant à la fois la modalité i de la variable A et la modalité j de la variable B.

	$B1$	$B2$	$\cdots$	Bj	$\cdots$	Bq	total
$A1$	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\cdot}$
$A2$	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2q}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ai	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ap	n_{p1}	n_{p2}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot q}$	$n_{\cdot\cdot}$

Remarques :
- les sommes marginales lignes sont $n_{i\cdot} = \sum_{j=1}^{q} n_{ij}$
- les sommes marginales colonnes sont $n_{\cdot j} = \sum_{i=1}^{p} n_{ij}$
- Les totaux des lignes sont identiques aux fréquences absolues issues de l'étude univariée de A.
- Les totaux des colonnes sont identiques aux fréquences absolues issues de l'étude univariée de B.
- Les sommes marginales sont liées entre elles par $n = n_{\cdot\cdot} = \sum_{j=1}^{q} n_{\cdot j} = \sum_{i=1}^{p} n_{i\cdot}$
- L'ordre d'entrée des variables dans la table de contingence n'a aucune importance. Mais on peut privilégier une des variables en constituant un tableau de profils associés aux lignes (respectivement aux colonnes).
- Le *tableau des profils* lignes (respectivement colonnes) est défini par les fréquences conditionnelles : $\frac{n_{ij}}{n_{i\cdot}}$ (respectivement $\frac{n_{ij}}{n_{\cdot j}}$). La somme de chaque ligne (respectivement colonnes) est alors ramenée à l'unité.

6.2 Le Chi-Deux de Contingence

Afin de mesurer l'intensité de la relation entre deux variables qualitatives, on calcule un paramètre statistique appelé *Chi-deux*, noté χ^2. Il s'agit de comparer les valeurs de la table de contingence observée avec les valeurs d'une table de contingence théorique.

Les données de la table de contingence théorique sont :
- les sommes marginales lignes identiques à celles de la table observée
- les sommes marginales colonnes identiques à celles de la table observée
- le nombre d'individus possédant à la fois le caractère i de la variable A et le caractère j de la variable B par :

$$\frac{n_{i\cdot} n_{\cdot j}}{n}$$

La valeur du Chi-Deux est maintenant définie par : $\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}}$

Le Chi-Deux peut s'énoncer de la manière suivante :

$$\chi^2 = \sum \frac{(\text{effectifs observés} - \text{effectifs théoriques})^2}{\text{effectifs théoriques}}$$

6.3 Test d'indépendance

L'hypothèse à tester concerne les deux variables qualitatives A et B observées sur un échantillon de n individus.

H_0 : " les deux variables sont indépendantes "

La statistique du Chi-Deux suit une loi du χ^2 à (p-1)(q-1) ddl.

Conditions d'application

1) Cas général

Pour appliquer un test du Chi-Deux, il faut que l'effectif total n soit grand et que les effectifs **théoriques** soient tous supérieurs à 5.

$$\frac{n_{i.}n_{.j}}{n} \geq 5$$

Si les effectifs théoriques sont voisins de 5, on peut appliquer la correction de continuité de Yates. En effet, les effectifs observés varient par saut d'une unité et leur approximation à la loi Normale qui est continue, engendre une sur-évaluation systématique du Chi-Deux.

$$\chi^2_c = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(\left|n_{ij} - \frac{n_{i.}n_{.j}}{n}\right| - 0.5\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Dans le cas où $p \geq 2$ et/ou $q \geq 2$, on peut regrouper (lorsque cela a un sens) les modalités des variables pour obtenir des conditions d'application correctes.

2) Cas particulier d'une table 2x2

Dans ce cas particulier, Cochran (1954) propose la règle suivante :

- si n<20, il faut appliquer le test exact de Fisher.

- si 20≤n<40 et si les effectifs théoriques sont supérieurs ou égaux à 5, on peut réaliser le test du Chi-Deux avec la correction de continuité de Yates.

- si 20≤n<40 et si certains effectifs théoriques sont inférieurs à 5 (mais ≥ 3), il faut :

soit augmenter la taille de l'échantillon

soit utiliser le test exact de Fisher

soit utiliser le test G qui est un peu plus robuste.

- si n≥40 et si les proportions ne sont voisines ni de 0, ni de 1, alors on peut réaliser le test classique du Chi-Deux (si non, test exact de Fisher).

Remarque : Les conditions d'application du Chi-Deux sont discutées depuis longtemps dans la littérature statistique. La plupart des livres les indique mais certains auteurs recommandent de ne pas les appliquer. Quoi qu'il en soit, il faut rester très vigilent quant à l'interprétation d'un test proche de la signification.

6.4 Exemple

Etudions le lien entre la pratique sportive individuelle (PSI) et le sexe (S) de $n = 300$ individus.

Table de contingence observée :

S \ PSI	Tennis	Athlétisme	Boxe	Ski	Total
Masculin	50	45	30	45	170
Féminin	55	30	0	45	130
Total	105	75	30	90	300

Tableau des profils lignes :

S \ PSI	Tennis	Athlétisme	Boxe	Ski
Masculin	0.294	0.265	0.176	0.265
Féminin	0.423	0.231	0	0.346

Tableau des profils colonnes :

S \ PSI	Tennis	Athlétisme	Boxe	Ski
Masculin	0.476	0.6	1	0.5
Féminin	0.524	0.4	0	0.5

Table de contingence théorique :

S \ PSI	Tennis	Athlétisme	Boxe	Ski	Total
Masculin	59.5	42.5	17	51	170
Féminin	45.5	32.5	13	39	130
Total	105	75	30	90	300

Le calcul du Chi-Deux de contingence donne la valeur $\chi^2 = 28.4099$.
Pour α=0.05 et $(p-1)(q-1)$=3 ddl, la valeur théorique du Chi-deux est 7.815.
Il existe un lien entre le sport pratiqué et le sexe.

Test sous R

```
tableau <- c(50, 55, 45, 30, 30, 0, 45, 45)
tableau <- matrix(tableau, nrow = 2)
chisq.test(tableau)
```

```
        Pearson's Chi-squared test

data:  tableau
X-squared = 28.4098, df = 3, p-value = 2.979e-06
```

6.5 Exercices

6.5.1 Discipline et latéralité

On connaît pour 119 étudiants en Activités Physiques et Sportives (APS) et 88 étudiants en Biologie la main préférentielle d'écriture. La table de contingence observée est :

	Droite	Gauche
APS	101	18
Biologie	81	7

Conclure.

6.5.2 Age et mode d'hébergement

Lors d'une enquête sur le tourisme et les pratiques de loisir en Ardèche [2], 2953 personnes ont été interrogées à l'aide d'un questionnaire bilingue français / anglais. 592 d'entre eux proviennent de la zone du moyen Vivarais. Peut-on mettre en évidence une différence d'âge entre campeurs et non campeurs ?
Les données se trouvent sur le site pédagogique dans le menu `Données` et le sous menu `Dossier de Fichiers` sous la dénomination `loisir.txt`.

7 Références Bibliographiques

Il existe en dehors du site et de l'excellent ouvrage de Stéphane Champely : La statistique vraiment appliquée au sport de nombreux ouvrages de référence. Certains sont très généraux ([1],[8]). D'autres sont plus accès sur des domaines d'application : plus médicale [6], plus biologie [9], plus sciences humaines [7]. Certains sont des grands classiques et retracent toute la statistique du plus simple au plus complexe ([3], [5], [4]). Enfin nous pouvons recommander deux ouvrages en ce qui concerne la statistique non paramétrique ([10], [11]).

Références

[1] Ceresta, editor. *Aide-mémoire pratique des techniques statistiques pour ingénieurs et techniciens supérieurs*. Centre d'Enseignement et de Recherche de Statistique Appliquée, Paris, 1991.

[2] P. Chazaud, A.B. Dufour, and B. Vignal. Vers uhe typologie des campeurs. l'exemple de l'ardèche. *Cahier Espaces*, 36 :61–68, 1994.

[3] P. Dagnelie. *Théorie et Méthodes Statistiques. Vol 1 : La statistique descriptive et les fondements de l'inférence statistique*. Presses Agronomiques de Gembloux, Gembloux, 1973.

[4] P. Dagnelie. *Analyse statistique à plusieurs variables*. Presses Agronomiques de Gembloux, Gembloux, 1975.

[5] P. Dagnelie. *Théorie et Méthodes Statistiques. Vol 2 : Les méthodes de l'inférence statistique*. Presses Agronomiques de Gembloux, Gembloux, 1975.

[6] S.A. Glantz. *Introduction aux biostatistiques*. McGraw-Hill, New-York, première édition, 1966 edition, 1998.

[7] J.P. Guilford and B. Fruchter. *Fundamental Statistics in Psychology and Education*. McGraw-Hill International Editions, Singapore, sixième édition edition, 1978.

[8] G. Saporta. *Probabilités, Analyse des Données et Statistiques*. Editions TECHNIP, Paris, 1990.

[9] B. Scherrer. *Biostatistique*. Gaëtan Morin, Québec, 1984.

[10] S. Siegel and N.J. Castellan. *Non Parametric Statistics for the Behavioral Sciences*. Statistics series. McGrawHill International, deuxième édition edition, 1989.

[11] P. Sprent. *Pratique des statistiques non paramétriques*. Editions INRA, Techniques et Pratiques, Paris, 1992.