# Raisonnements dans l'analyse de données expérimentales en sciences de l'éducation

## Compléments

Luc-Olivier Pochon

**Luc-Olivier Pochon**

# Raisonnements dans l'analyse de données expérimentales en sciences de l'éducation

## Troisième partie

Version 2c, décembre 2013

# Troisième partie : Compléments techniques

> *La science est un dialogue avec la nature … Toute prise de mesure, préalable à la création de connaissance, présuppose la possibilité d'être affecté par le monde, que ce soit nous qui soyons affectés ou nos instruments. (Ilia Prigogine, 1996:177).*

## Présentation

L'ouvrage[1] dont ce document est la troisième partie constitue une invitation à la relecture des différents types de « raisonnements » qui interviennent dans l'analyse de données. La présentation est adaptée aux étudiants en sciences de l'éducation, public particulier dans la mesure où il est à la fois utilisateur des outils mais aussi intéressé professionnellement à leur existence en tant qu'objets de connaissance et d'enseignement.

Cette troisième partie est constituée de compléments qui détaillent des aspects techniques dépassant les besoins courants, mais qui permettent toutefois d'accéder à certains fondements omis la plupart du temps. Sous une forme unifiée, ils s'attachent à présenter les idées à la base de certaines procédures en tenant compte des difficultés les plus souvent rencontrées. Quelques développements techniques formels sont destinés à un public plus mathématicien qui voudrait entrer dans le sujet.

Le premier complément (Chapitre III.1) propose, en guise de rappel, une présentation schématique des concepts, théories et outils traités.

Les trois chapitres suivants présentent des distributions théoriques. Le chapitre III.2 en rappelle la définition et donne la liste des plus communes. Le chapitre III.3 s'attache plus précisément à la distribution liée à la courbe logistique et le chapitre suivant (Chapitre III.4) à celle du « $\chi^2$ ».

[1] Pochon, L.-O. (2012). Raisonnements dans l'analyse de données expérimentales en sciences de l'éducation. Paris : L'Harmattan, Figures de l'interaction.

Le chapitre III.5 fait le tour des logiciels de statistiques avec notamment la présentation de deux outils mis à disposition par l'Institut de psychologie et éducation de l'Université de Neuchâtel, ANASTAT et une bibliothèque de fonctions pour **R**. Sur la base de l'analyse d'un questionnaire, le chapitre suivant (Chapitre III.6) montre quelques exemples d'utilisation de **R**.

Les quatre chapitres suivants introduisent, à partir d'exemples, les modèles d'analyses globales couramment utilisés ou mentionnés : analyse factorielle en composantes principales (ACP) (Chapitre III.7) ; analyse factorielle des correspondances (AFC) (Chapitre III.8) ; analyse log-linéaire (Chapitre III.9) ; analyse de variance (Chapitre III.10).

Les trois chapitres suivants sont consacrés au coefficient *S* de Kendall. Après une comparaison entre distribution exacte et approchée (Chapitre III.11), l'étude d'une interaction à l'aide de ce coefficient (Chapitre III.12) est proposée. Quelques propriétés formelles de ce coefficient sont ensuite présentées (Chapitre III.13).

Les trois chapitres suivants concernent les systèmes complexes. Le chapitre III.14 présente la suite logistique. Le chapitre suivant (Chapitre III.15) donne des exemples de constructions de fractales. Finalement, un développement concernant le calcul de l'entropie constitue le chapitre III.16.

Le chapitre III.17 constitue une introduction au vocabulaire de la théorie des graphes.

Finalement, de dernier chapitre (Chapitre III.18) propose une description des fonctions de la bibliothèque IPErad pour **R** qui se trouvait sur le site de l'Institut de psychologie et éducation de l'université de Neuchâtel[1].

[1] Que l'on trouvera actuellement, sous la rubrique analyse de données sur la page: http://abord-ch.org/cours/welcome.html

## Chapitre III.1. le champ de l'analyse de données

Le schéma de la figure ci-contre situe le processus de l'analyse de données en science de l'éducation (SED) dans un environnement scientifique général tel qu'il est envisagé dans l'ouvrage.

Dans ce schéma, on distingue tout d'abord les grandes catégories (en orange) qui concernent l'activité scientifique : épistémologie, la méthode scientifique en général, les domaines scientifiques dont les théories liées constituent une particularisation.

La deuxième catégorie (en brun) concerne les « outils », matériels ou intellectuels. On y trouve les théories de la confirmation et de la validation, les outils informatiques, divers schémas de validation et un ensemble de techniques liées à la logique et au calcul des probabilités.

Les éléments de ce dernier ensemble constituent des champs scientifiques pour eux-mêmes. A noter que la psychologie et les sciences de l'éducation (en rouge) entretiennent un double rapport aux éléments de cette catégorie. Ce sont à la fois des outils, mais aussi des champs d'applications et d'études (aspects cognitifs, pédagogiques, etc.).

Une autre catégorie est constituée des démarches (en vert) : expérimentation, observation, « large scale assessment », etc. Ces démarches conduisent d'hypothèses théoriques à divers types de prises de décisions (en bleu). Ces démarches intègrent le traitement de données (en blanc).

Les domaines des sciences (faits, lois, démarches)
Dépend de/génèrent
La méthode scientifique
Influence/Alimente
Epistémologie
Se formalise en
Théories de la validation
Théories de la confirmation
Sciences & techniques
Activité
Instrumen-te/talise
Outils (informatiques)
Statistique et probabilité
Théorie de l'information
Systémique (sc. de la complexité)
Logique
Fait partie / contient
Aspects psycho-cognitifs
"Théorie" ou "question" en psychologie ou en SED
Génère
Hypothèses théoriques
Permet l'élaboration
Modifie
Implique
Schémas de validation
Mise à l'épreuve
Traitement
Hypothèses aux.
Hypothèses observables
Plan expérimental
Petits échantillons
Simulation
Large scale assessment
Dispositif (CTT, IRT)
Données
Mise en oeuvre
Résultats
Conduit à
Interprétation, décision, action (IDA)
Obtenue par
Test
Questionnaire

## Chapitre III.2. Les distributions de probabilités ou lois de probabilités

Il est possible de distinguer les lois discrètes et les lois continues à densité. Après un exemple de chaque type, quelques distributions fréquemment utilisées sont présentées.

### Lois discrètes

Une loi discrète est donnée par la probabilité de chaque événement. Il s'agit souvent d'une formule algébrique permettant de calculer $p(x)$ ou une table de valeurs, ou encore la probabilité cumulée.

Les probabilités sur l'espace des événements $U$ peuvent être associées aux valeurs prises par une variable aléatoire $X$.

$$p(x) = p(X = x) = p(E) \text{ avec } E = \{e \in U : X(e) = x\}$$

*Exemple (loi binomiale)*

Evénement considéré : cinq lancers d'une pièce de monnaie « idéale », avec :

- $U$ = {PPPPP, FPPPP, PFPPP, …, FFFFF} ($U$ possède 32 éléments) ;

- $X$ = nombre de piles (P).

Le tableau III.2.1 donne les valeurs de la probabilité de chaque événement et les probabilités cumulées.

| X | Probabilité | Probabilité cumulée |
|---|---|---|
| 0 | 1/32 | 1/32 |
| 1 | 5/32 | 6/32 |
| 2 | 10/32 | 16/32 |
| 3 | 10/32 | 26/32 |
| 4 | 5/32 | 31/32 |
| 5 | 1/32 | 1 |

**tableau III.2.1**. Table des valeurs d'une distribution binomiale

### Lois à densité

Les probabilités sur $U$ peuvent être associées aux valeurs prises par $X$ sur un intervalle de valeurs (cas continu).

$$p(a<X<b) = p(E) \text{ avec } E = \{e \in U : a<X(e)<b\}$$

Il y a deux façons complémentaires de donner pratiquement la loi : par la densité *f* de probabilité ou par la fonction de répartition (*cumulative distribution function* - CDF).

La fonction *f*, densité de probabilité, permet de calculer la probabilité sur un intervalle :

$$p(a < X < b) = \int_a^b f(x)d\mu(x)$$

La fonction de répartition est dans ce cas $F(x) = p(-\infty < X < x)$. On a alors, $p(a<X<b) = F(b) - F(a)$. Si la fonction de répartition *F* est donnée d'abord, la densité de probabilité *f* est la dérivée de *F*.

*Exemple (la loi normale)*

Les graphes de f et F dans le cas de la distribution normale centrée réduite sont donnés dans la figure III.2.1.

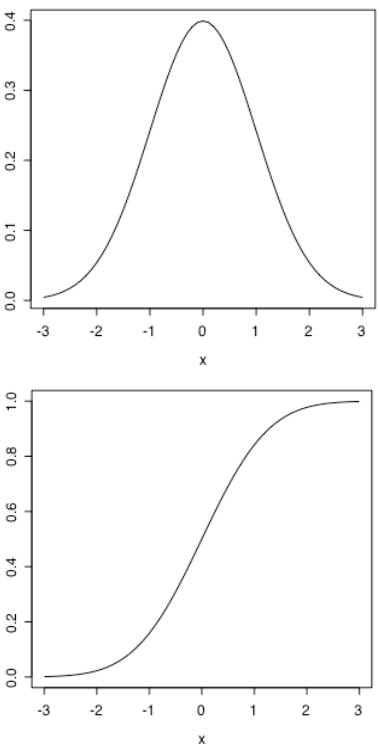



**fig III.2.1**. Densité de probabilité et fonction de répartition de la loi normale

**Exemples de lois de probabilités théoriques**

*Loi binomiale*

La distribution ou loi de probabilité la plus simple est la distribution (discrète) binomiale. La loi de paramètres $n$ et $p$ est liée à la répétition d'une épreuve aléatoire à deux issues $E$ et $E'$, de probabilités $p$ et $q$=1-$p$. La loi $X$ est le nombre d'apparitions de $E$ lors de $n$ répétitions de l'épreuve.

La loi est donnée par : $p(X=i)=\frac{n!}{i!(n-i)!}p_i(1-p_i)^{n-i}$ La moyenne de cette distribution est $E(X)=np$ et la variance $var(X)=npq$ (écart-type : $\sqrt{npq}$ ).

*Loi multinomiale*[1]

Cette loi généralise la loi binomiale. La loi de paramètres $n$, $m$, $(p_i)_{i=1,\ldots,m}$ avec $\sum_{i=1}^{m} p_i=1$ est donnée par la répétition à $n$ reprises d'une épreuve aléatoire à $m$ issues $E_i$, chacune de probabilité $p_i$. En notant $N_i$ le nombre d'occurrences de $E_i$, on a :

$$p(N_1=n_1,\ldots,N_m=n_m)=\frac{n!}{n_1!\ldots n_m!}p_1^{n_1}\ldots p_m^{n_m} \text{ avec } \sum_{i=1}^{m} n_i=n \ .$$

Les $N_i$ sont des variables binomiales. Cela se conçoit directement à partir de la définition ou par calcul. On a :

$p(N_1=n_1)=\sum_{n_2+\ldots+n_m} p(N_1=n_1,\ldots,N_m=n_m)$ et donc :

$$p(N_1=n_1)=\frac{n!}{n_1!(n-n_1)!}p_1^{n_1}\sum_{n_2+\ldots+n_m}\frac{(n-n_1)!}{n_2!\ldots n_m!}p_2^{n_2}\ldots p_m^{n_m}$$

La somme constituant le troisième facteur vaut justement par généralisation de la formule du binôme : $(p_2+\ldots+p_m)^{n-n_1}$ .

On a :

$E(N_i)=np_i$ ; $var(N_i)=np_i(1-p_i)$ ;

$cov(N_i,N_j)=-n\,p_i\,p_j$ ;

$$corr(N_i,N_j)=-\sqrt{\frac{p_i\,p_j}{(1-p_i)(1-p_j)}} \ .$$

[1] http://fr.wikipedia.org/wiki/Loi_multinomiale

*Loi de Poisson*[1]

Cette loi, discrète et infinie, prolonge la loi binomiale au cas infini.

Son application prototypique est le suivant. On considère un événement pouvant se produire une ou plusieurs fois sur un intervalle de temps $T$, en moyenne $\lambda$ fois. On appelle $X$ la variable aléatoire déterminant le nombre de fois où l'événement se produit dans la période $T$. $X$ prend des valeurs entières : 0, 1, 2, ...

Cette variable aléatoire de paramètre $\lambda$ suit une loi de probabilité définie par : $p(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$ . Sa moyenne et son écart-type valent $\lambda$.

*Loi normale*[2]

C'est la loi que suit une variable qui est la somme d'une multitude de paramètres aléatoires. Elle permet d'approcher de nombreuses autres lois notamment la loi binomiale lorsque le nombre d'épreuves est grand (théorème de Moivre-Laplace). Le graphe de la densité de probabilité (distribution) est la fameuse courbe en cloche qui répond à l'expression analytique :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Les paramètres $\mu$ et $\sigma$ sont respectivement la moyenne et l'écart type de la distribution. La loi normale centrée réduite est normalisée par $\mu = 0$ et $\sigma = 1$. La fonction de répartition associée est notée $\Phi$.

*Loi du $\chi^2$ (chi-2)*

La loi du $\chi^2$ à $d$ degrés de liberté correspond à la distribution de la somme de $d$ variables, chacune suivant une loi normale centrée réduite (voir chapitre III.4).

[1] http://fr.wikipedia.org/wiki/Loi_de_Poisson
[2] http://fr.wikipedia.org/wiki/Loi_normale

*Loi du F (Fisher-Snedecor)*

Cette loi utilisée dans l'analyse de variance donne la distribution d'une variable résultant du rapport de deux loi du $\chi^2$.

*Loi logistique*

Cette loi est donnée par la fonction de répartition $F(x) = \frac{1}{1+e^{-\alpha(x-\beta)}}$. La densité de probabilité $f(x)$ associée se calcule facilement en dérivant $F(x)$ :

$$f(x) = F'(x) = \frac{\alpha e^{-\alpha(x-\beta)}}{(1+e^{-\alpha(x-\beta)})^2}$$

Cette distribution est de moyenne β et d'écart-type $\frac{\pi}{\alpha\sqrt{3}}$ (voir chapitre III.3).

*Loi de Cauchy*

Cette loi met en évidence le fait qu'il s'agit de composer avec notre intuition et subir en quelque sorte les lois du monde III de Popper.

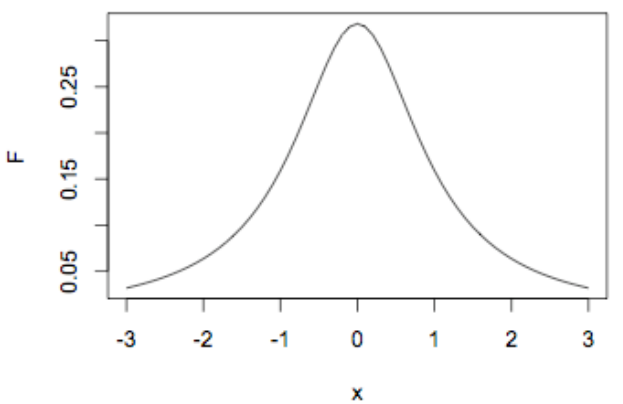



**fig III.2.2**. Distribution de Cauchy

La loi de Cauchy réduite est donnée par la fonction de répartition :

$$F(x) = \frac{1}{\pi(1 + x^2)}.$$

La courbe correspondante (figure III.2.2) à l'apparence d'une courbe en cloche. La médiane est, selon la définition empirique, la valeur 0.

Par contre, il n'est pas possible de calculer la moyenne et la variance. Le calcul formel conduit à des valeurs infinies.

Pour des expériences conduisant à des résultats qui suivent une loi de Cauchy, la loi des grands nombres n'est pas valable. En d'autres termes, la suite des moyennes obtenues en augmentant le nombre de valeurs expérimentales ne tend pas vers une valeur fixe.

Si $X$ et $Y$ sont deux variables indépendantes de distribution normale réduite, alors $X / Y$ suit une loi de Cauchy réduite (Bastien, 1960).

La loi de Cauchy générale est obtenue par translation et dilatation.

## Chapitre III.3. Evidence, logit et Logit

La vraisemblance (odds) d'un événement E est le rapport $p$(E)/$p$(E'). Si $p = p$(E), ce rapport vaut $p$/(1-$p$). Irving John Good ou Alan Turing (la paternité de l'idée n'est pas clairement établie) ont proposé une manière de manier plus facilement les probabilités en introduisant la quantité (log-odds) : $10 \log \frac{p}{1-p}$ . Par son apparentement avec l'expression de lois physiques, l'unité est le *décibel* [db]. Cette valeur est proportionnelle à $\ln \frac{p}{1-p}$ dont l'unité est le *logit*. Il est parfois signalé que *logit* serait l'abréviation de *log-odds-unit*. Il existe par ailleurs une fonction $\mathrm{Logit}(p) = \ln \frac{p}{1-p}$ dont la genèse est autre.

### Les fonctions Probit et Logit

La fonction Probit est la fonction inverse de la distribution cumulative associée à la distribution normale centrée réduite.

$$z_0 = \mathrm{Probit}(p) \Leftrightarrow \Phi(z_0) = p(z < z_0) = p$$

La distribution normale est donnée par une densité de probabilité qui forme une courbe en cloche. La distribution cumulative ou fonction de répartition associée $\Phi$ est une courbe en S (figures III.3.1a et III.3.1b).

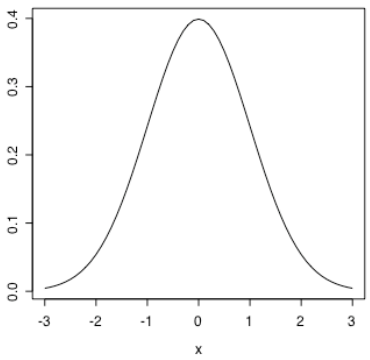



**fig III.3.1a**. Loi normale : densité de probabilité

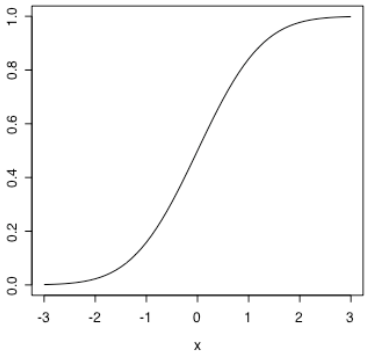



**fig III.3.1b**. Loi normale : distribution cumulative Φ

La fonction logistique est donnée par : $p = \frac{1}{1+e^{-\alpha(x-\beta)}}$. Elle varie entre 0 et 1. Son graphe est une courbe en S. Considérée comme une fonction cumulative d'une distribution, cette dernière est également une courbe en cloche (figures III.3.2a et III.3.2b).

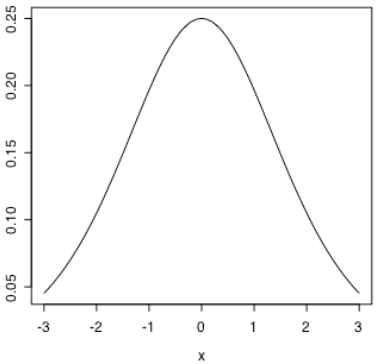



**fig III.3.2a**. Loi logistique : densité de probabilité ($\alpha = 1$ et $\beta = 0$)

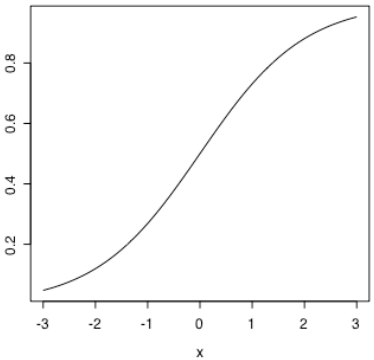



**fig III.3.2b**. Loi logistique : distribution cumulative ($\alpha = 1$ et $\beta = 0$)

**Le modèle de Rasch**

Dans ce modèle, la « capacité » des personnes à répondre à des items est indiquée dans la même unité que celle de la difficulté des items. Le modèle à un paramètre est donné par :

$$p = \frac{1}{1+e^{\beta - x}} \text{ ou } p = \frac{e^{x-\beta}}{1+e^{x-\beta}}$$

où : $\beta$ est la difficulté de l'item, $p$ est la probabilité pour une personne de capacité $x$ de répondre correctement à l'item. $\beta$ et $x$ s'expriment en *logits*.

$$x - \beta = \ln \frac{p}{1-p} \; [logits]$$

A une différence de capacités correspond un rapport de vraisemblance. Ainsi, si entre deux personnes, la différence de capacité est de 1 *logit*, le rapport de vraisemblance de réussite au même item (ou à des items de même difficulté) vaut $e \approx 2.72$.

Pour un item de difficulté moyenne ($\beta = 0$), on a les correspondances suivantes :

| Capacité [logit] | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Vraisemblance | 0.05 | 0.14 | 0.37 | 1 | 2.72 | 7.39 | 20.08 |
| Probabilité de réussite | 0.05 | 0.12 | 0.270 | 0.5 | 0.73 | 0.88 | 0.95 |

De même la même manière, on obtient le tableau qui exprime pour un élève moyen ($x = 0$) la correspondance entre difficulté d'un item et la probabilité de réussite. Cette probabilité de réussite est aussi le taux de réussite dans une population distribuée symétriquement autour des élèves « moyens ».

| Difficulté [logit] | 3 | 2 | 1 | 0 | -1 | -2 | -3 |
|---|---|---|---|---|---|---|---|
| Vraisemblance | 0.05 | 0.14 | 0.37 | 1 | 2.72 | 7.39 | 20.08 |
| Probabilité de réussite | 0.05 | 0.12 | 0.270 | 0.5 | 0.73 | 0.88 | 0.95 |

**Modèle à deux paramètres**

Il est donné par la distribution cumulative :

$$S(x) = \frac{e^{\alpha(x-\beta)}}{1+e^{\alpha(x-\beta)}}.$$

La distribution est de moyenne β et d'écart-type $\pi/\alpha\sqrt{3}$. α est le pouvoir discriminant de l'item. Pour une valeur $p$ de $S(x)$, on a : $x - \beta = \frac{1}{\alpha}\ln\frac{p}{1-p}$ .

**Pour réfléchir**

*A propos de l'usage du modèle de Rasch dans la théorie de réponse à l'item (IRT)*

Le modèle à deux paramètres est donné par : $x - \beta = \frac{1}{\alpha}\ln\frac{p}{1-p}$ ou par $p = \frac{1}{1+e^{-\alpha(x-\beta)}}$ avec $x$ « capacité » de l'élève, $p$ probabilité pour un élève de capacité $x$ de réussir un item de difficulté β et de coefficient de discrimination α.

1) Calculez la probabilité pour un élève moyen ($x$=0), de réussir un item de coefficient $\alpha = 2$, en fonction de son niveau de difficulté β.

2) En supposant une population d'élèves « équilibrée » par rapport à l'élève moyen, donnez une estimation du pourcen-

tage de réussite à des items présentant les différents degrés de difficulté de -3 [logit] à 3 [logit].

On pourra utiliser les formules simplifiées :

- La vraisemblance $v = \frac{p}{1-p}$ en fonction de la difficulté β est donnée par : $v = e^{-2\beta}$.

- La probabilité $p$ en fonction de la vraisemblance : $p = \frac{v}{v+1}$.

| Difficulté β [logit] | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Vraisemblance $p/(1-p)$ | | | | 1 | | | |
| Probabilité de réussite $p$ | | | | 0.5 | | | |

3) Pour un item moyen (β = 0) de valeur discriminante α = 1, la distribution cumulative (fonction de répartition) de la réussite des élèves en fonction de leur "capacité" $x$ est donnée par : $x = \ln\frac{p}{1-p}$. Montrez que la distribution correspondante est de moyenne 0 et d'écart type d'environ 1.81.

4) Complétez le tableau dont la dernière ligne est liée à la distribution de la capacité selon une échelle de moyenne 500 et d'écart-type 100.

| Capacité x [logit] | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Vraisemblance de réussite $p/(1-p)$ | 0.05 | | | | | | |
| Probabilité de réussite $p$ | 0.05 | | | | | | |
| Capacité selon la nouvelle échelle | 334 | | | | | | |

Quelle sera la probabilité pour un élève de capacité 510 de réussir un item de difficulté moyenne?

Indication : La capacité selon la nouvelle échelle se calcule à l'aide la formule: $100x/1.81 + 500$ (transformation d'une échelle par une application affine).

## Chapitre III.4. A propos du « $\chi^2$ »

Le $\chi^2$, *chi-2* ou *khi2* est un symbole qui concerne à la fois une distribution, une mesure et un test statistique.

### $\chi^2$ comme distribution

La loi normale centrée réduite est définie par la densité de probabilité :

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

La fonction de répartition correspondante est : $\Phi(z_o) = \int_{-\infty}^{z_0} \varphi(z)dz$ qui donne la probabilité, pour la variable $z$, d'être comprise entre $-\infty$ et $z_0$ .

Il est parfois utile de considérer la variable $z^2$ (qui sera notée $\chi^2$) toujours positive et dont la densité de probabilité peut facilement être déduite par changement de variable de la précédente[1] :

$$f(t) = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t}{2}}}{\sqrt{t}}$$

La situation précédente se généralise. On peut considérer la variable :

$$\chi_k{}^2 = z_1{}^2 + z_2{}^2 + \ldots + z_k{}^2$$

où $z_i$ est une variable de loi normale centrée réduite. $\chi_k{}^2$ a comme densité de probabilité : $f(t) = \frac{1}{2^{k/2}\Gamma(k/2)} t^{\frac{k}{2}-1} e^{-\frac{t}{2}}$ où $k$ est le degré de liberté et $\Gamma$ la fonction « Gamma » d'Euler qui prolonge la fonction « factorielle » ($\Gamma(n+1) = n!$ pour $n$ entier).

### $\chi^2$ comme mesure

Dans une expérience d'issues possibles $E_i$, $i$ variant de 1 à $k$, on considère lors de $n$ répétitions de l'expérience le nombre d'occurrences, $O_i$, de chacune des issues $E_i$, et l'effectif

[1] t est toujours positif. Il devrait être remplacé par $\chi^2$ ce qui crée des confusions.

théorique [1] $T_i$. Pour mesurer l'écart entre la théorie et la pratique, on[2] considère les différences élevées au carré (pour supprimer un nombre négatif éventuel), puis on divise par la valeur attendue pour rendre les écarts relatifs à cette valeur. On obtient le coefficient donné par :

$$\Delta = \frac{(O_1 - T_1)^2}{T_1} + \ldots + \frac{(O_k - T_k)^2}{T_k}$$

Sous certaines hypothèses sur la distribution des écarts entre théorie et pratique, notamment lorsque les issues $E_i$ sont liées à une loi normale, le coefficient $\Delta$ suit une loi du $\chi^2$ avec $k$ degrés de liberté[3].

*Un usage basé sur les simples probabilités*

En appliquant ce procédé à une expérience dichotomique où les deux issues possibles $E_1$ et $E_2$ ont des probabilités d'occurrence de $p$ et $q$ ($p+q = 1$), cette formule devient en notant $X$ le nombre d'occurrences de $E_1$[4] :

$$\frac{(X - np)^2}{Np} + \frac{(n - X - nq)^2}{Nq} =$$

$$\frac{q(X - np)^2 + p(n - X - nq)^2}{npq} =$$

$$\frac{(X - np)^2}{npq}$$

$X$ est une variable aléatoire obéissant à une loi binomiale (moyenne $np$, variance $npq$). Si $n$ est grand, la loi binomiale peut être approchée[5] par loi normale de moyenne $np$ et de variance $npq$.

---

[1] Cet effectif théorique peut dépendre d'une distribution théorique, loi normale, de Poisson, etc.

[2] Le procédé a été proposé par Pearson principalement pour vérifier la normalité de certains résultats.

[3] Dans certaines présentations, c'est la probabilité associée qui est prise comme mesure plutôt que la valeur de $\chi^2$. Cette procédure a l'avantage de rendre la mesure indépendante du degré de liberté. L'hypothèse de normalité est souvent omise d'autant plus qu'elle est valable dans tous les cas basés sur des tirages aléatoires.

[4] Par rapport à la formule générale : $O_1 = X$, $O_2 = n - X$, $T_1 = np$, $T_2 = nq$ .

[5] On notera le changement de point de vue. Précédemment, l'hypothèse de normalité s'appliquait à l'événement. Ici, il s'agit d'un outil technique

D'où $z = \frac{X - np}{\sqrt{npq}}$ suit une loi normale centrée réduite. Et donc le coefficient précédent, noté également $\chi^2$, suit la loi du $\chi^2$ à 1 degré de liberté (notée : $\chi_1{}^2$). On a : $\chi^2 = \chi_1{}^2 = \frac{(X - np)^2}{npq}$

Ces considérations se généralisent au cas multinomiaux. Dans le cas d'issues à k valeurs, on peut écrire :

$$\chi^2_{k-1} = \frac{(O_1 - T_1)^2}{T_1} + \ldots + \frac{(O_k - T_k)^2}{T_k} = \sum_i \frac{(O_i - T_i)^2}{T_i}.$$

Le nombre de degrés de liberté est *k*-1 puisque la somme des $O_i$ (égale à celle des $T_i$) est fixée à *n*.

En utilisant les fréquences $q_i = O_i / n$ et $p_i = T_i / n$, on a :

$$\chi^2_{k-1} = n\left(\frac{(q_1 - p_1)^2}{p_1} + \ldots + \frac{(q_k - p_k)^2}{p_k}\right) = \sum_i \frac{(q_i - p_i)^2}{p_i}$$

$$\text{avec} \sum_i p_i = \sum_i q_i = 1$$

**Le test du $\chi^2$**

Lorsqu'il s'agit de tester la vraisemblance d'une série d'issues d'une expérience aléatoire, dans le cas du rejet d'une hypothèse $H_0$ de pur hasard, il est possible de substituer sans grande perte de précision (d'autant plus que *n* est grand) la loi multinomiale régissant les fréquences d'apparitions des différentes issues possibles par la loi du $\chi^2$. Celle-ci se trouve tabulée dans les manuels de statistiques ou programmée dans les packages statistiques ou dans les principaux tableurs.

**Exemple numérique**

On considère comme expérience le lancer d'une pièce de monnaie. L'expérience est répétée à $n = 10$ reprises. Les valeurs observées (*O*) et théoriques (*T*) sont données dans le tableau III.4.1.

---

simplificateur. Le calcul exact serait possible, comme dans le test exact de Fisher. Dans un cas, le hasard s'exprime de façon « continue » (c'est la distribution gaussienne), dans l'autre cas, il s'exprime directement par des probabilités.

| | P | F | total |
|---|---|---|---|
| *O* | 3 | 7 | 10 |
| *T* | 5 | 5 | 10 |

**tableau III.4.1**. Effectifs observés et théoriques dans 10 parties de « pile ou face »

On trouve facilement à partir des effectifs ou des fréquences la valeur du $\chi^2$ :

$$\chi^2 = \frac{(3-5)^2}{5} + \frac{(7-5)^2}{5} = \frac{8}{5} = 1,6$$

A noter que la probabilité donnée par la distribution théorique avec l'hypothèse d'une pièce équilibrée ($\chi_1^2$ d'une valeur supérieure à 1.6) vaut 0.45. L'hypothèse que la pièce est biaisée ne peut pas être remise en cause.

### Usage du $\chi^2$ pour l'étude de la liaison de deux variables nominales[1]

Dans le cas précédent, le coefficient $\chi^2$ était utilisé pour mesurer la distance entre deux distributions « quelconques » (test d'homogénéité) ou entre une distribution empirique et son équivalent théorique (test de conformité). Ce procédé peut aussi servir à mesurer l'indépendance (test d'indépendance) entre deux types familles d'issues E = ($E_i$) et E' = ($E'_j$) d'une même expérience (ou deux caractères d'une population), *i* compris entre 1 et *k*, *j* compris entre 1 et *m*.

En effet, ce cas peut se ramener à une différence de deux distributions :

- la distribution des effectifs constatée d'apparition de ($E_i$ et $E'_j$) : $O_{ij}$ ;

- la distribution théorique sous l'hypothèse d'indépendance des deux caractères : $T_{ij}$ .

L'hypothèse d'indépendance implique que $T_{ij}$ est à la fois proportionnel au nombre d'occurrences de $E_i$ et à celui de $E'_j$. Sa valeur se calcule à l'aide du quotient : $T_{ij} = O_i\ O'_j\ /\ n$

[1] Dans la littérature, on trouve aussi les termes « qualitatives » ou « catégoriels ».

où $O_i$ (resp. $O'_i$) est le nombre d'occurrences de $E_i$ (resp. $E'_i$).

| | $E'_1$ | $E'_2$ | $E'_3$ | Total |
|---|---|---|---|---|
| $E_1$ | 2 (4) | 2 (2) | 6 (4) | $O_1$=10 |
| $E_2$ | 8 (6) | 3 (3) | 4 (6) | $O_2$=15 |
| Total | $O'_1$=10 | $O'_2$=5 | $O'_3$=10 | $n$=25 |

**tableau III.4.2**. Table de contingence construite pour deux caractères E et E' (entre parenthèse les effectifs théoriques)

Le $\chi^2$ se calcule comme précédemment avec $d = (k\text{-}1)(m\text{-}1)$ comme degré de liberté (c'est le nombre de coefficients du tableau que l'on peut choisir librement).

Ces informations peuvent être représentées dans une table de contingence (tableau III.4.2).

**Calcul à partir des fréquences**

Lorsque l'on travaille à partir des fréquences, les notations suivantes sont souvent utilisées :

- Fréquence empirique de ($E_i$ et $E'_j$) : $p_{ij}$.

- Fréquence empirique et théorique de $E_i$. Elles sont égales par convention : $p_{\bullet i} = \sum_j p_{ij}$ .

- Fréquence empirique et théorique de $E'_j$. Elles sont égales par convention : $p_{j\bullet} = \sum_i p_{ij}$ .

- Fréquence théorique de ($E_i$ et $E'_j$) sous l'hypothèse d'indépendance : $p_{\bullet i} p_{j\bullet}$ .

Avec ces notations : $\chi^2 = n \sum_{ij} \frac{(p_{ij} - p_{i\bullet} p_{\bullet j})^2}{p_{i\bullet} p_{\bullet j}}$.

**Mesure de l'association**

Sur la base du coefficient $\chi^2$, on définit des coefficients qui donne le degré d'association (sorte de corrélation) entre les deux caractères[1] :

[1] Une autre mesure d'association est le coefficient « kappa » ($\kappa$) de Cohen qui donne une mesure d'accord. Il n'est pas basé sur le $\chi^2$.

Coefficient de contingence : $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

Coefficient de contingence relatif : $C_r = \frac{C}{C_{\max}}$ avec $C_{\max} = \sqrt{k' - 1 / k'}$ où $k'$ est le minimum entre le nombre de lignes et le nombre de colonnes.

Phi de Cramér[1] : $\Phi_c = \sqrt{\frac{\chi^2}{n(k'-1)}}$

Ce coefficient se trouve également sous le nom de V de Cramér. Pour un tableau 2x2, ce coefficient $\Phi = \sqrt{\frac{\chi^2}{n}}$ était utilisé par Pearson comme coefficient de corrélation.

**Exemple numérique**

Avec les valeurs du tableau III.4.2, le tableau III.4.3 donne les résultats en fréquence.

| | $E'_1$ | $E'_2$ | $E'_3$ | Total |
|---|---|---|---|---|
| $E_1$ | .08(.16) | .08(.08) | .24(.16) | $p_{1.}$= .40 |
| $E_2$ | .32(.24) | .12(.12) | .16(.24) | $p_{2.}$= .60 |
| Total | $p_{.1}$=.40 | $p_{.2}$=.20 | $p_{.3}$=.40 | 1 |

**tableau III.4.3**.Table de contingence avec les fréquences

On vérifie que les valeurs obtenues pour le $\chi^2$ à partir des effectifs et des fréquences sont les mêmes :

Effectifs : $\chi^2 = \frac{2^2}{4} + 0 + \frac{2^2}{4} + \frac{2^2}{6} + 0 + \frac{2^2}{6} = 3.333$

Fréquences : $\chi^2 = 25\left(\frac{0.08^2}{0.16} + 0 + \frac{0.08^2}{0.16} + \frac{0.08^2}{0.24} + 0 + \frac{0.08^2}{0.24}\right) = 3.333$

[1] Harald Cramér (1893-1985), mathématicien et statisticien suédois. A ne pas confondre avec Gabriel Cramer (1704-1752), mathématicien suisse né à Genève, « inventeur » de la règle de Cramer pour résoudre les systèmes d'équations (mais qui a également tâté des probabilité en proposant une solution au paradoxe de St Petersburg).

En prenant la distribution $\chi^2$ avec 2 degrés de liberté, on trouve : $p(\chi^2 > 3.333) = 0.19$. Dans ce cas, on peut difficilement rejeter le fait que les deux caractères sont indépendants. C'est aussi ce que permettent d'entrevoir les valeurs « un peu élevées » des coefficients d'association :

- Coefficient de contingence : $C = \sqrt{\frac{3.333}{3.333 + 25}} = 0.343$ ; $C_{\max} = \sqrt{1/2} = 0.707$ ; $C/C_{\max} = 0.49$.

- Phi de Cramer : $\Phi_c = \sqrt{\frac{3.333}{25(2-1)}} = 0.365$.

**L'approche par la théorie de l'information**

Le rapport de similitude entre deux distributions de probabilité $p = (p_i)$ et $q = (q_j)$ est donné par (Kullback, 1968) : $\sum_1^n q_i \log_2\left(\frac{q_i}{p_i}\right)$.

C'est la moyenne (selon $q$) de la quantité d'information par observation en faveur de $q$ « contre » $p$. On peut faire de même en échangeant les rôles de $p$ et $q$ et considérer : $\sum_1^n p_i \log_2\left(\frac{p_i}{q_i}\right)$

Ces deux valeurs sont positives ou nulles. Leur somme est la *divergence* entre les deux distribution :

$$J(p,q) = J(q,p) = \sum_1^n (q_i - p_i) \log_2\left(\frac{q_i}{p_i}\right).$$

Lorsque le rapport est proche de 1, on peut écrire : $\ln\left(\frac{q_i}{p_i}\right) = \ln\left(1 + \frac{q_i}{p_i} - 1\right) \cong \frac{q_i}{p_i} - 1$ et donc[1] :

[1] On utilise le début du développement de ln ($x$+1) = 1+$x$-$R$, où $R$ est inférieur à $x^2/2$. Cette approximation est robuste vu la compensation qui s'opère entre le facteur formé d'une soustraction et le facteur logarithme.

$$J(p,q) = \sum_1^n (q_i - p_i) \log_2 \left( \frac{q_i}{p_i} \right) \cong \frac{1}{\ln 2} \sum_1^n (q_i - p_i) \left( \frac{q_i}{p_i} - 1 \right) =$$

$$\frac{1}{\ln 2} \sum_1^n \frac{(q_i - p_i)^2}{p_i} = \frac{1}{n \ln 2} \chi^2$$

*Calcul de la divergence*

Pour les données du tableau A4.1, on obtient :

$$J(p,q) = \sum_1^n (q_i - p_i) \log_2 \left( \frac{q_i}{p_i} \right) = -0.2 \log_2 \frac{0.3}{0.5} + 0.2 \log_2 \frac{0.7}{0.5} =$$

$$0.2 \log_2 \frac{0.7}{0.3} = 0.244$$

On vérifie que la valeur fournie à partir de $\chi^2$ est proche :

$$J(p,q) \cong \frac{1}{n \ln 2} \chi_1^{\,2} = \frac{1}{10 \ln 2} 1.6 = 0.23$$

**Liaison de deux variables nominales : approche par la théorie de l'information**

L'information moyenne apportée par la connaissance de la répartition de ($E_i$) vaut avec les notations introduites page 23 : $I_m(E) = -\sum_i p_{i\bullet} \log_2 p_{i\bullet}$

L'information moyenne apportée par la connaissance de la répartition de (E'$_i$) vaut : $I_m(E') = -\sum_j p_{\bullet j} \log_2 p_{\bullet j}$

L'information moyenne apportée par la connaissance simultanée des deux répartitions vaut : $I_m(E \ et \ E') = -\sum_{ij} p_{ij} \log_2 p_{ij}$

---

Pour le calcul de $\sum_1^n q_i log_2 \left( \frac{q_i}{p_i} \right)$, une meilleure approximation permet d'obtenir $\sum_1^n q_i log_2 \left( \frac{q_i}{p_i} \right) \cong \frac{1}{2n \ln 2} \chi^2$.

L'information mutuelle entre E et E' est la quantité positive et symétrique par rapport à E et E' donnée par[1] :

$$I_m(E,E') = I_m(E) + I_m(E') - I_m(E \ \ et \ \ E') = \sum_{ij} p_{ij} \log_2 \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}}$$

On peut vérifier que lorsque le quotient est proche de 1, le $\chi^2$ fournit une approximation[2] de ce coefficient[3] (pour simplifier l'écriture, on remplace $p_{ij}$ par $o_i$ et $p_{\cdot i} p_{j\cdot}$ par $t_i$) :

$$\sum_i o_i \ln \frac{o_i}{t_i} \cong \sum_i (o_i - t_i + t_i)\left( \frac{o_i - t_i}{t_i} - \frac{1}{2}\left( \frac{o_i - t_i}{t_i} \right)^2 \right) =$$

$$\sum_i \frac{(o_i - t_i)^2}{t_i} - \frac{1}{2}\sum_i \frac{(o_i - t_i)^3}{t_i^{\,2}} + \sum_i (o_i - t_i) - \frac{1}{2}\sum_i \frac{(o_i - t_i)^2}{t_i} =$$

$$\frac{1}{2}\sum_i \frac{(o_i - t_i)^2}{t_i} - \frac{1}{2}\sum_i \frac{(o_i - t_i)^3}{t_i^{\,2}} \cong \frac{1}{2}\sum_i \frac{(o_i - t_i)^2}{t_i}$$

Finalement : $I_m(E,E') \cong \dfrac{1}{\ln 2}\displaystyle\sum_{ij} \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}} = \frac{1}{2n \ln 2}\chi^2$

Dans le cas du tableau A4.3, on obtient : $I_m(E,E') = 0.0996$

La clause de proximité des valeurs empiriques et théoriques n'étant pas respectées, l'approximation de l'information mutuelle entre E et E' donnée par le $\chi^2$ n'est pas bonne tout en restant acceptable, en effet : $\dfrac{1}{2n \ln 2}\chi^2 = 0.0962$

**Le coefficient $G^2$**

Un test fréquemment mis en œuvre (Croux, 2005 ; Kah & Pruvot, 2003), le test du rapport de vraisemblance, utilise la statistique (avec les notations introduites page 23) :

$$G^2 = 2\sum_i O_i \ln \frac{O_i}{T_i} = 2n \sum_i q_i \ln \frac{q_i}{p_i}$$

---

[1] C'est la moyenne (selon la distribution effective) de quantité d'information par observation en faveur de l'indépendance « contre » la distribution effective.

[2] On utilise $\ln(1+x) = x - x^2/2 + R$ avec $R$ inférieur à $x^3/3$.

[3] Lebart & Fénelon (1973) omettent le facteur 1/2.

Dans le cas de l'étude de la liaison entre deux variables nominales, les calculs menés précédemment permettent d'établir la relation :

$$G^2 = 2n \sum_{ij} p_{ij} \ln \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} = 2n \ln 2 \quad I(E,E') \cong \chi^2$$

Lorsque deux distributions théoriques ($T_i$) et ($S_i$) sont en concurrence, on peut considérer :

$$G^2(S \mid T) = G^2(T) - G^2(S) = 2 \sum_i O_i \ln \frac{S_i}{T_i}$$

## Chapitre III. 5. Les progiciels de statistique

Les progiciels dédiés à l'analyse de données peuvent être répartis en six catégories.

- Tableurs : tous les tableurs courants possèdent des fonctions statistiques. Des feuilles de calcul et macros dédiées à certains usages existent que ce soit sous un mode mutualisé ou commercial.

- Les progiciels « historiques »[1] de statistique qui proposent une interface de tableur (par exemple : SPSS, STATISTICA). Ces systèmes permettent de conduire de nombreuses analyses à travers des boîtes de dialogue. Un langage de scripts permet d'automatiser des procédures.

- Les progiciels spécialement conçus comme outil didactique (par exemple : ANASTAT).

- Les environnements de calcul statistique (par exemple : **R**, **S**). Ces environnements travaillent en lignes de commande dans une « console » (des librairies sont à disposition qui permettent d'assurer un fonctionnement par « menus »). Des librairies (packages) spécialisées enrichissent une palette déjà bien fournie d'outils statistiques de base.

- Les environnements de calcul numérique généraux (Mathlab) et certains environnements de calcul symbolique (Mathematica) possèdent des fonctions statistiques qui peuvent être enrichies par l'ajout de modules.

- Certains environnements sont spécialisés dans des domaines bien délimités. C'est le cas notamment de WEKA (Waikato Environment for Knowledge Analysis)[2] qui présente un nombre important d'algorithmes dans le domaine de la classification automatique.

---

[1] On désignera par ce terme les systèmes issus des « packages » statistiques disponibles sur les premiers ordinateurs. Nous citons ici les systèmes les plus souvent utilisés en sciences humaines. La liste pourrait comprendre encore, les historiques SAS, P-STAT, etc. et les nouveaux venus, SPHINX notamment qui allie à la fois des outils statistiques des services pour réaliser des enquêtes en ligne.

[2] http://en.wikipedia.org/wiki/Weka_(machine_learning) (consulté : janvier 2010)

Il faut encore citer les systèmes d'analyse textuelle tels que Alceste[1], QDA Miner[2], Tropes[3], etc[4]. Ces outils d'analyse textuelle se trouvent souvent sous la forme d'un couplage d'outils d'aide à la catégorisation et à l'indexation et d'outils d'analyses ou de représentations standard.

Quelques exemples d'utilisation vont être donnés pour les quatre premières catégories.

**Le tableur**

A titre d'exemple, la figure III.5.1 présente le calcul d'un *chi-2* réalisé de façon élémentaire. Pour une utilisation plus intensive du tableur, il est possible d'utiliser des « add-on », notamment XLSTAT[5] qui semble être l'outil d'analyse de données et de statistiques pour EXCEL le plus complet et le plus utilisé.

chi-2.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | brun | vert | gris | total |
| 2 | fille | 2 | 5 | 8 | 15 |
| 3 | garçon | 8 | 10 | 2 | 20 |
| 4 | total | 10 | 15 | 10 | 35 |
| 5 | | | | | |
| 6 | | | | | |
| 7 | | 4.2857143 | 6.4285714 | 4.2857143 | 15 |
| 8 | | 5.7142857 | 8.5714286 | 5.7142857 | 20 |
| 9 | | 10 | 15 | 10 | 35 |
| 10 | | | | | |
| 11 | | | | | |
| 12 | | 1.2190476 | 0.3174603 | 3.2190476 | |
| 13 | | 0.9142857 | 0.2380952 | 2.4142857 | |
| 14 | | | | | |
| 15 | ddl = | 2 | | C = | 0.4382928 |
| 16 | chi-2 = | 8.3222222 | | PHI = | 0.4876246 |
| 17 | p = | 0.0155902 | | | |
| 18 | | | | | |

**fig III.5.1**. Calcul du *chi-2* à l'aide d'un tableur

[1] http://www.image-zafar.com/index_alceste.htm (consulté : juin 2011).
[2] http://www.provalisresearch.com/QDAMiner/QDAMinerDesc.html (consulté : juin 2011).
[3] http://www.tropes.fr/ (consulté : juin 2011)
[4] Les vénérables Nu*dist, The Ethnograph ne semblent plus être disponibles.
[5] http://www.xlstat.com (consulté : octobre 2008)

Dans cette feuille de calcul, les formules suivantes sont introduites :

- dans E2 : `SOMME(B2 :D2)` copiée dans `E3,E4,E7,E8,E9`;
- dans B4 : `SOMME(B2 :B3)` copiée dans `C2,D2,B9,C9,D9` ;
- dans B7 : `B$4*$E2/$E$4` copiée dans `C7,D7,B8,C8,D8` ;
- dans B12 : `(B2-B7)^2/B7` copiée dans `B12 :D13` ;
- dans B16 : `SOMME(B12:B12:D13:D13)` ;
- dans B17 : `LOI.KHIDEUX(B16;B15)` ;
- dans E15 : `RACINE(B16/(B16+E4))` ;
- dans E16 : `RACINE(B16/E4)`.

**Les progiciels historiques**

SPSS[1] est le logiciel phare des sciences sociales, actuellement en concurrence avec d'autres systèmes, comme STATISTICA[2]., issus des domaines techniques et industriels. Comme EXCEL pour le gestionnaire, SPSS se pose parfois comme un élément identitaire du chercheur en sciences sociales.

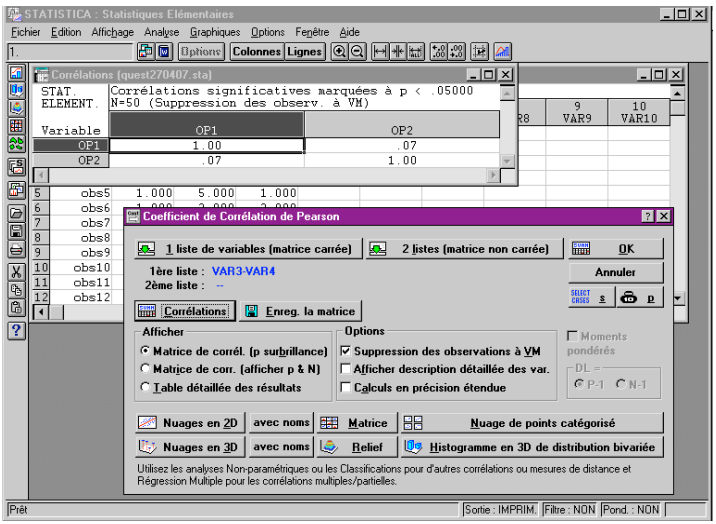



**fig III.5.2**. Boîte de dialogue de STATISTICA liée au calcul d'un coefficient de corrélation

[1] http://www.spss.com/fr/ (consulté : août 2008).
[2] http://www.statsoft.fr/ (consulté : août 2008).

L'utilisateur interagit avec ces logiciels à travers des boîtes de dialogue dont la figure III.5.2 présente un exemple.

**Les environnements statistiques :** R

**R** (**R** Development Core Team, 2009) est un système interactif par lignes de commande qui bénéficie d'une communauté[1] d'utilisateurs parmi les chercheurs en éducation. Des manuels à l'intention spécifique des psychologues se trouvent sur Internet (Pallier & Lalanne, 2005). Il est possible de hiérarchiser les compétences nécessaires à son utilisation.

*Niveau I*

La figure III.5.3a présente les données et la figure III.5.3b les manipulations de base (niveau I) rendues aisées par la commande `attach()`. Cette commande rend les variables (statistiques) directement disponibles à l'appel de leur nom. Dans cet exemple, les données de base consistent en une cinquantaine d'enregistrements et quatre variables : `sexe` (1 ou 2), `age` (1, 2 ou 3), `op1`, `op2`, deux degrés d'adhésion à deux opinions sous la forme d'une échelle de Lickert (1 à 5). Les manipulations effectuées sont les suivantes : calcul de la moyenne de `op1`, établissement du tableau croisé `sexe x op1` (en son enregistrement dans l'objet `matable`), réalisation de deux tests sur ce tableau croisé, chi-2 et test de Jonkheere. Le premier test est disponible dans les outils de base du système. Le deuxième nécessite une bibliothèque supplémentaire.

```
> quest <- read.table( "dataRAD.txt")
> quest
"iden" "sexe" "age""op1" "op2"
"obs1"  1      1     3     3
"obs2"  1      1     4     5
"obs3"  1      1     4     4
"obs4"  1      1     4     5
"obs5"  1      1     5     5
"obs6"  1      1     3     3
"obs7"  1      1     2     4
…
> attach(quest)
```

**fig III.5.3a**. Chargement des données **R**

[1] http://fr.wikipedia.org/wiki/R_(logiciel)

Le test *t de Student* a été ajouté pour montrer l'opération de restriction du nombre de cas à partir d'un test sur la valeur d'une variable.

```
> mean(op1)
[1] 2.89
> matable <- table(sexe,op1)
> matable
op1
sexe 1  2  3  4  5
  1  2  6  5 12  2
  2  1 16  5  4  1

# test du chi-2

> chisq.test(matable)
data: matable
X-squared = 9.2121, df = 4,
p-value = 0.05601

# test de Jonkheere (S Kendall)
# chargement d'une bibliothèque

> source("IPErad")
> S.test(matable)
S = -258, Correction = 12.75,
Sc = -245.25, variance = 12028.25,
ecart-type = 109.7, z = -2.24,
p-value = 0.013

# test de Student
>t.test(op1[sexe==1],op1[sexe==2])
…
```

**fig III.5.3b**. Des manipulations élémentaires effectuées avec **R**

*Niveaux II à IV*

Le deuxième niveau de maîtrise concerne l'usage des fonctions vectorisées et le mode de sélection des cas en fonction de valeur de variables. Par exemple :

`> op1[1:25]` (sélection des 25 premières valeurs de la variable `op1`);

`> op1[op1!=0] <- 1` (toutes les valeurs non nulles de `op1` sont mises à 1).

On entre dans le niveau III lorsqu'il faut construire ou déconstruire des objets et donc comprendre différents types de structures informatiques (vecteurs, listes, dataframes, etc.).

Finalement le dernier niveau concerne la programmation de nouvelles fonctions qui, pour être efficace, demande des habitudes non usuelles en programmation classique comme, par exemple, le remplacement de boucles par des produits scalaires ou extérieurs.

**L'environnement ANASTAT**

Le logiciel ANASTAT[1] (anciennement ANAPROT) est un projet développé pour l'Institut de psychologie pour différents usages, notamment pour permettre d'effectuer des tests non- paramétriques (à l'époque de son développement, ceux-ci ne figuraient pas dans la panoplie des packages historiques), effectuer des calculs exacts (paradigme sans population) et traiter des séquences d'observations. Développé primitivement dans un environnement Lisp (Pochon, 1991) pour assurer une bonne interactivité (fouille de données), il a été converti en Prolog dans une version qui a été finalement compilée et « recouverte » d'un système de menu, ce qui lui a permis de subsister à travers les âges de l'informatique personnelle.

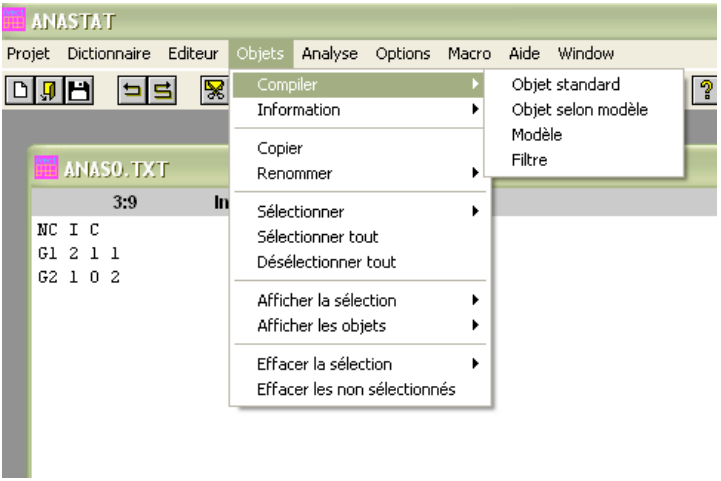



**fig III.5.4**. Compilation des données avec ANASTAT

[1] http://www2.unine.ch/ipe/home/liens/utilitaires_statistiques (consulté : août 2012)

Quelques figures permettent de présenter une manipulation liée à la recherche d'une « p-value » exacte[1]. La figure III.5.4 montre les données qui vont être « compilées » selon un modèle (ici un objet standard de type tableau croisé) pour constituer un objet `tab1`. La manipulation suivante va sélectionner cet objet et générer les différents *patterns* qui partagent des propriétés communes avec lui. On obtient les objets `tab1;i` (*i*=1,…,7).

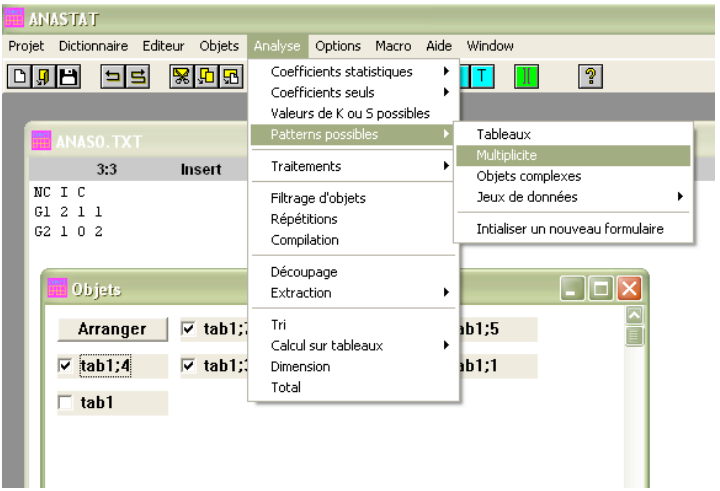



**fig III.5.5**. Recherche des multiplicités de chacun des *patterns* possibles

La manipulation suivante (figure III.5.5) permet d'obtenir la multiplicité[2] de chacun de ces tableaux. La figure III.5.6 montre le résultat obtenu qui permettra de juger de la rareté du *pattern* initial (observé) par rapport aux cas possibles. En associant un coefficient statistique à ce calcul (par exemple

[1] Cette opération poursuit le même but didactique que la procédure proposée par Roditi (2009). Toutefois, notre approche se base sur la définition théorique d'une probabilité (nombre de cas favorables rapporté au nombre de cas possibles). Dans l'autre cas, la définition est fréquentiste. Cette approche pose le problème de situer deux distributions (celle des tirages et celle du *chi-2*) par contre elle ne demande pas, comme l'approche théorique, de limiter les effectifs pour des raisons d'explosion combinatoire. Dutarte (2006) propose une troisième voie où la simulation est directement effectuée sur le phénomène brut et non à travers une mise en forme (tableau) ou un coefficient (*chi-2*).

[2] Ce coefficient est lié au traitement des ex-aequo dans les procédures basées sur l'usage des rangs.

le S de Kendall), on en obtient la distribution théorique exacte.

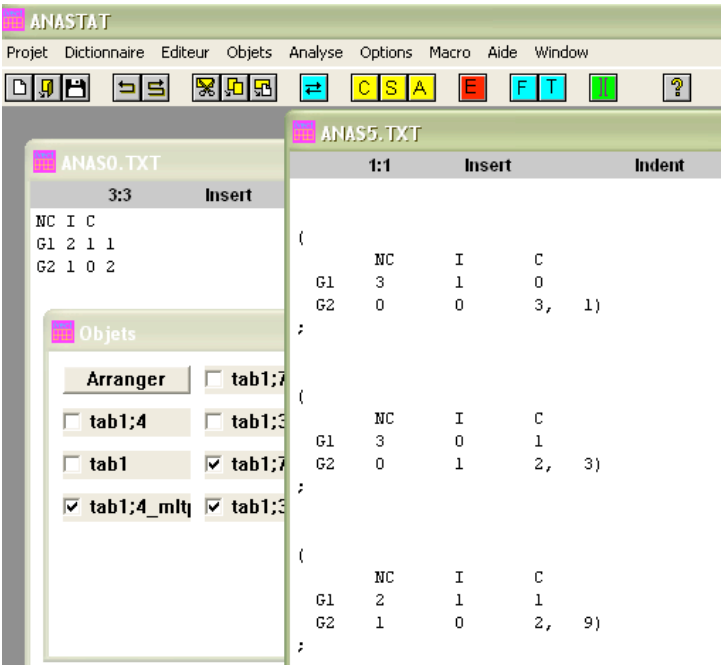



**fig III.5.6**. Les *patterns* possibles accompagnés de leur multiplicité (tab1;i_mltp, i=1,…,7)

## Chapitre III.6. Analyse d'un questionnaire

*- L'entrée en matière est très importante : les premières questions indiquent à l'enquêté le type de sujet dont il va être question.*
*- Faire très attention aux transitions entre les questions et éviter les ruptures brusques.*
*- Regrouper les questions par thèmes.*
*- ...*[1]

Ce chapitre prend l'étude du questionnaire pour passer en revue un certain nombre de concepts, de techniques et d'outils.

### Types de questions d'un questionnaire

On distingue des caractéristiques des questions selon leur type et leur forme. Le type indique si la question concerne une information factuelle (âge, par exemple), une opinion générale, ou une opinion à propos du répondant lui-même.

Du point de vue forme, il est usuel de considérer les questions à choix multiples (selon plusieurs modalités possibles), les questions ouvertes (numériques, textuelles courtes et longues) et les échelles (principalement Lickert, Thurstone ou Guttman). Parmi les questions à choix multiples, on distingue parfois les réponses dichotomiques (oui, non) des autres.

Il faut distinguer la forme de la présentation. Ainsi une « case à cocher » ou une alternative (les boutons « radio » oui et non) sont deux présentations de la même forme.

### Contexte de l'étude

Le contexte est composé d'un contexte général (universel) et d'un contexte particulier. Le premier concerne la population étudiée, la théorie dans laquelle l'étude s'inscrit avec, le cas échéant, les hypothèses théoriques associées, des théories auxiliaires sur le mode de questionnement. Le contexte particulier est formé de l'échantillon ou des groupes interrogés et des questions particulières choisies.

---

[1] http://www.animafac.net/article.php3?id_article=150 (Consulté : juin 2007)

**Portée des résultats**

L'étude avec un questionnaire peut poursuivre plusieurs buts, notamment les trois suivants :

- Rechercher des « paramètres » d'une population sur la base d'échantillons, par exemple : le pourcentage de la population ayant une certaine opinion (ou opinion moyen) sur un sujet donné, la mesure de l'effet d'un caractère (sexe, âge) sur une opinion (ce qui revient à comparer deux populations). La passation du questionnaire livre des résultats chiffrés dont l'analyse peut faire intervenir un jugement statistique selon les deux paradigmes, échantillon / population ou *population-free*.

- Explorer les caractéristiques de groupes. Dans ce cas, il s'agit de statistiques purement descriptives qui vont être appelées à la rescousse.

- Mesurer la différence entre des groupes selon divers critères. Dans ce cas, le jugement statistique peut servir à une validation externe (généralisation des observations) ou interne (détermination de la précision permise par l'instrument de mesure).

**Les types d'analyse**

Pour illustrer ce point, on suppose avoir à disposition un questionnaire avec deux questions factuelles (sexe : femme, homme et âge : jeune, moyen, âgé) et deux questions pour mesurer une opinion sur une échelle de Lickert (de 1 à 5).

```
> quest = read.table("dataQuest.txt")
> attach(quest)
> quest
"iden" "sexe" "age" "op1" "op2"
"obs1" 1 1 3 3
"obs2" 1 1 4 5
"obs3" 1 1 4 4
"obs4" 1 1 4 5
"obs5" 1 1 5 5
"obs6" 1 1 3 3
"obs7" 1 1 4 4
```

**fig III.6.1**. Les données de base

Sauf mention du contraire, on utilisera la syntaxe de **R** pour présenter les exemples numériques (voir chapitres III.5 et

III.18). La figure III.6.1 présente le chargement des données et le début de la liste.

On peut regrouper les analyse en cinq catégories :

*1. Dénombrements simples*

Il s'agit par exemple de calculer le pourcentage ou le nombre d'individus ayant répondu à chaque modalité d'une réponse à choix multiples ou d'une échelle (figure III.6.2). Dans le cas d'une échelle, la distribution exacte est à préférer à la moyenne si cela est possible.

```
# effectif
> table(op1)
op1
 1  2   3  4  5
 2 22  10 16  3
# fréquence
> table(op1)/ sum(table(op1))
op1
    1     2     3     4     5
0.056 0.41 0.19 0.30 0.06
> mean(op1)
[1] 2.89
```

**fig III.6.2**. Dénombrements des effectifs, calcul de la moyenne, fréquences

*2. Comparaison entre des sous-groupes*

La figure III.6.3 illustre la création d'un tableau croisé, nommé *matable*, à partir duquel on effectue le test du *chi-2* (figure III.6.4) et le test de Jonkheere (figures III.6.5 et III.6.6).

```
> table(sexe,op1)
    op1
sexe  1  2  3  4  5
   1  2  6  5 12  2
   2  1 16  5  4  1
```

**fig III.6.3**. Tableau croisé

```
> matable=table(sexe,op1)
> chisq.test(matable)
data: matable
X-squared = 9.2121, df = 4, p-value = 0.05601
```

**fig III.6.4**. Test du chi-2

Pour ce dernier, il a fallu charger une « bibliothèque ». La figure suivante (figure III.6.7) présente le test du t de Stu-

dent (paramétrique) réalisé à partir des données brutes auquel peut-être substitué celui de Wilcoxon[1] qui est équivalent dans ce cas à celui de Jonkheere.

```
> source("IPErad.r")
> S.test(matable)
S = -258, corr = 12.75, Sc = -245.25
variance = 12028.25, ecart-type = 109.7
z = -2.24, p-value = 0.013
```

**fig III.6.5**. Test de Jonkheere avec chargement d'une bibliothèque

Sans indication complémentaire, le test de Wilcoxon et le test du *t de Student* sont bilatéraux (considère la différence sans s'intéresser au sens de cette différence). Le test de Jonkheere est unilatéral.

```
> S.test(table(age,op2))
S = -110, corr = 1, Sc = -109
variance = 15014.59, ecart-type = 122.5
z = -0.89, p-value = 0.19
```

**fig III.6.6**. test de Jonkheere sur trois groupes

```
> t.test(op1[sexe==1],op1[sexe==2])
Welch Two Sample t-test
data: op1[sexe==1] and op1[sexe==2]
t = 2.3746, df = 50.355, p-value = 0.02142
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval: 0.1028557
1.2304777
sample estimates:
 mean of x mean of y
 3.222222  2.555556
```

**fig III.6.7**. La différence d'opinion (op1) selon le sexe testée à l'aide du *t .de Student*

*3. Comparaison entre questions*

La figure III.6.8 propose une comparaison entre deux questions à l'aide d'un *t de Student* unilatéral avec l'option d'appariement.

[1] Commande :
> `wilcox.test(op1[sexe==1],op1[sexe==2])`

```
> t.test(op1,op2,paired=T,alternative="g")
Paired t-test
data: op1 and op2
t = 0.348, df = 53, p-value = 0.3646
alternative hypothesis: true difference in means
is greater than 0
95 percent confidence interval: -0.2823171
Inf
sample estimates: mean of the differences
0.07407407
```

**fig III.6.8**. Cas apparié avec R (test unilatéral)

Cette comparaison peut également s'effectuer en considérant le tableau croisant deux questions (figure III.6.9).

```
> tab12 <- table(op1,op2)
> tab12
   op2
op1 1 2 3 4 5
  1 0 0 2 1 0
  2 5 8 6 1 2
  3 0 1 5 3 1
  4 3 5 2 3 3
  5 1 0 1 0 1
```

**fig III.6.9**. Tableau croisé construit sur les deux questions q1 et q2

Selon l'hypothèse un *Chi-2* ou le calcul du coefficient *S de Kendall* (avec la *p-value* correspondante) permet de juger du degré de similarité des réponses aux deux questions[1]. Une autre alternative est présentée ci-dessous.

*4. Regroupement des questions*

Pour amorcer le regroupement des questions (figure III.6.10), il peut être procédé à des calculs de corrélations. Des analyses multivariées (analyse factorielle, régression multiple) peuvent systématiser cette recherche.

```
> cor.test(op1,op2)
Pearson's product-moment correlation
data: op1 and op2
t = 0.7967, df = 52, p-value = 0.4293
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval: -0.1627346
0.3667868
sample estimates: cor 0.1098109
```

**fig III.6.10**. Corrélation

[1] Ici Chi-2 = 17.769, df = 16, p-value = 0.3376

*5. Analyse des questions ouvertes*

Différents outils, plus ou moins automatisés, peuvent être utilisés pour l'analyse des questions ouvertes. Comme signalé dans le chapitre III.5, ces systèmes sont tout d'abord constitués d'outils d'aide à la catégorisation et à l'indexation du corpus de textes. Une fois les catégories constituées et dénombrées, des fonctions standard peuvent prendre le relais pour analyser ou pour représenter les données numériques obtenues.

## Chapitre III.7. Exemple d'analyse factorielle en composantes principales (ACP)

Cette annexe est divisée en trois parties. La première partie décrit de façon un peu technique le raisonnement à la base de cette analyse. La deuxième partie illustre, sur un exemple simple, les principales notions et raisonnement qui interviennent dans cette technique. Elle montre également comment l'analyse factorielle peut être utilisée dans l'enseignement de « tous les jours » dans l'analyse des tests d'évaluation en opérant une classification des questions (en distinguant, par exemple, celles dont la difficulté est liée au traitement et celles où la compréhension de l'énoncé représente la difficulté principale). Elle offre également la possibilité de mettre en évidence des profils d'élèves. La dernière partie présente les données intermédiaires produites par le système dont la lecture n'est pas forcément d'une utilité immédiate.

### Présentation synthétique de la méthode

Cette présentation est à l'intention des lecteurs familiers avec la notion d'espace vectoriel muni d'un produit scalaire. On suppose pour fixer les idées que *n* élèves répondent à *m* questions chacune notée de 1 à 6. Pour présenter les calculs mis en œuvre, deux voies principales sont possibles[1]. La première considère l'espace vectoriel des sujets dont un repère est indexé par les questions. La figure 14 du chapitre I.4, illustre ce cas qui, dans son développement, va utiliser un résultat concernant les formes hermitiennes (recherche d'une base orthogonale pour deux formes bilinéaires, voir Doneddu, 1986:230). La deuxième est plus intuitive et sera utilisée ici. Elle met en œuvre l'espace des tests.

Dans cette version, on considère un espace vectoriel E de dimension *n* muni d'une base orthonormée indicée par les élèves (on notera $\bullet \mid\mid \bullet \mid\mid$ le produit scalaire et la norme habituelle). L'ensemble des résultats à la question numéro *i* représente un vecteur $X_i$ dans cet espace. Ces vecteurs sont

[1] Pochon, L.-O. (1981). Analyse factorielle : notations et procédés. Document IRDP non publié.

ensuite « réduits », on soustrait tout d'abord à chaque composante la moyenne des composantes $\mu(X_i)$. On obtient le vecteur $X_i - \mu(X_i)U$ où $U$ est le vecteur dont toutes les composantes sont 1. On normalise le vecteur résultant en le divisant par l'écart type de $X_i$ : $\sigma(X_i)$. Finalement, on considère :

$$M_i = \frac{X_i - \mu(X_i)U}{\sigma(X_i)} \text{ avec :}$$

$$\mu(M_i) = 0 \ ; \ \sigma(M_i) = 1 \ ; \ \|M_i\|^2 = n \ ;$$

$$\frac{M_i \bullet M_j}{n} = corr(X_i, X_j) = corr(M_i, M_j)$$

La matrice $M$ est la matrice où les colonnes sont les vecteurs[1] $M_i$.

La matrice $\frac{1}{n} M^t M$ est la matrice des corrélations. Le coefficient $c_{ij}$ de cette matrice est la corrélation de Bravais-Pearson de $X_i$ et $X_j$ , $corr(X_i, X_j)$, mais aussi de $M_i$ et $M_j$ .

La matrice $C = \frac{1}{n} MM^t$ n'a pas d'interprétation simple. Par contre, elle est de dimension $nxn$, ses vecteurs colonnes sont des combinaisons linaires des vecteurs $M_i$ et ses valeurs propres sont les mêmes que celles de la matrice des corrélations (qui existent, les matrices étant symétriques).

Le principe de la méthode est de trouver une direction telle que la variance des projections des vecteurs $M_i$ sur cette direction soit maximale, puis de répéter le processus.

En formule, on cherche $F$ tel que : $I = \sum corr(F, M_i)^2$ maximum ; $\mu(F) = 0$ ; $\sigma(F) = 1$ .

Avec ce vecteur F, on aura: $\frac{1}{n} F^t M_i = corr(F, M_i)$ et donc :

$$I = \left(\frac{1}{n} F^t M\right)\left(\frac{1}{n} F^t M\right)^t = \frac{1}{n^2} F^t M M^t F = \frac{1}{n} F^t C F$$

$C$ est une matrice symétrique. Elle peut donc se décomposer de la manière suivante (Doneddu, 1986:228) : $C = PDP^{-1}$ où

[1] Que l'on peut supposer linéairement indépendants si le nombre d'élèves est supérieur au nombre de questions.

$P$ est une matrice orthogonale (donc son inverse est égale à sa transposée) et $D$ une matrice diagonale. On a donc :

$$I = \frac{1}{n}(P^{-1}F)^t D(P^{-1}F)$$

On prend $V$ le vecteur dont toutes les composantes sont nulles sauf la composante correspondant à la valeur propre maximale $\lambda = \lambda_{max}$ qui vaut $\sqrt{n}$. Il est facile de vérifier que :

$I = \frac{1}{n}V^t DV = \lambda_{max}$ est maximal avec la contrainte $\|V\|^2 = n$.

$F=PV$ est le vecteur candidat. On vérifiera que $\mu(F)=0$ et $\sigma(F)=1$. La formule ci-dessus dira alors que $\sum corr(F,M_i)^2$ est maximum (et vaut $\lambda_{max}$) (propriété III.7.1).

$F$ est vecteur propre de $C$ (puisque $\lambda V=DV$), il est donc combinaison linéaire des colonnes de $C$. Par conséquent, il est combinaison linéaire des vecteurs $M_i$ : $F = \sum \alpha_i M_i$ . Cela implique que $\mu(F)=0$. De plus, la norme de $V$ vaut $\sqrt{n}$, donc celle de $F$ aussi. Cela implique que $\sigma(F)=1$. On a encore :

- $corr(F,M_i) = \lambda\alpha_i$ (propriété III.7.2) ;

- $n = F^t F = \sum \alpha_i n \times corr(F,M_i)$

- et donc $1 = \lambda \sum \alpha_i^2$ ou $\sum \alpha_i^2 = \frac{1}{\lambda}$

En effet : $\lambda F = CF = (\frac{1}{n}MM^t)F = \frac{1}{n}M(M^t F)$. Comme $M^t F$ est le vecteur colonne de composante $n \times corr(M_i,F)$ on a :

$$\sum \lambda\alpha_i M_i = \lambda F = \frac{1}{n}M\begin{pmatrix} n \times corr(F,M_1) \\ \cdots \\ n \times corr(F,M_m) \end{pmatrix} = \sum corr(F,M_i)M_i$$

On en déduit la première propriété : $corr(F,M_i) = \lambda\alpha_i$. La deuxième propriété découle immédiatement de la première.

**Une analyse factorielle « ordinaire »**

*La situation*

Les données concernent un cours d'informatique suivi dans une école professionnelle par 27 étudiants. Les résultats obtenus tout au long de l'année sont réunis dans un tableau

de 27 lignes (une par étudiant, ce sont les « observations ») et 19 colonnes (les « notes », ce sont les « variables ») dont voici la description:

- TEST1: note obtenue à un contrôle de connaissance (brèves questions ouvertes);

- TEST2: note obtenue à un contrôle de connaissance (QCM)

- NOTE_THE: moyenne ajustée des deux résultats précédents. L'information est redondante, mais a été ajoutée pour des questions pratiques et sert également de point de repère pour le niveau général.

Ces notes varient de façon classique de 1 à 6. Les 7 notes suivantes sont des appréciations concernant des aptitudes plus générales. Elles peuvent prendre les valeurs 0, 0.5, 1 :

- CONF: degré de conformité des rapports rendus aux standards proposés ;

- JUSTE: degré de justesse des informations figurant dans les rapports ;

- INTERET: intérêt du contenu général des rapports ;

- CLARETE: clarté dans la présentation ;

- P_EN_MAIN: intérêt des exemples mentionnés pour une prise en main ;

- AGRESSIV: « agressivité » montrée dans la recherche d'information ;

- EXPOSE: appréciation concernant l'exposé oral d'un travail personnel ;

La valeur suivante est MOY_TP qui est une moyenne ajustée (note de 1 à 6) des 7 résultats précédents. La même remarque que pour NOTE_THE s'applique ici.

Les 7 valeurs suivantes représentent la notation de l'épreuve d'examen constituée de problèmes plus complexes sensés mesurer des compétences (entre parenthèses l'étendue des points pouvant être attribués), soit EX1 (0-6), EX2 (0-6), EX3 (0-6), EX4 (0-2), EX5 (0-4), EX6 (0-6), EX7 (0-6). La dernière valeur (MOY_EXAM) est la moyenne ajustée des 7 notes précédentes (cette note varie entre 1 et 6).

Par exemple les valeurs des variables attribuées au premier étudiant, ROA, sont : 4.5, 3.0, 3.8, 1.0, 1.0, 0.5, 1.0, 0.5, 0.5, 0.5, 4.57, 6.0, 3.5, 2.0, 2.0, 2.0, 5.0, 2.0, 4.5.

*L'analyse*

Une analyse en composantes principales est effectuée. Quatre facteurs sont demandés qui représentent respectivement 31%, 14%, 10% et 9 % de la variance (variation) totale. En tout, 65% de la variance (variation totale) est expliquée. Le tiers de la variance restera inexpliqué.

Le pas suivant est de trouver une interprétation des facteurs. On considère pour cela le tableau (figure III.7.1) présentant les corrélations des anciennes variables par rapport aux facteurs (qui sont aussi les coordonnées des variables réduites dans la base des facteurs).

| FACTOR ANALYSIS | Extraction: Principal components (Marked loadings are > .700000) | | | |
|---|---|---|---|---|
| Variable | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
| TEST1 | .760936 | .200086 | −.157736 | −.339531 |
| TEST2 | .618369 | −.279992 | −.277983 | −.469496 |
| NOTE_THE | .794680 | −.040639 | −.260073 | −.465668 |
| CONF | .231397 | −.025305 | .492916 | −.334833 |
| JUSTE | .252667 | −.316905 | −.142941 | .603274 |
| INTERET | .073705 | −.672066 | .121617 | −.308335 |
| CLARETE | .572101 | .376346 | −.335855 | .333423 |
| P_EN_MAI | .355473 | −.386432 | −.305890 | .099714 |
| AGRESSIV | .469545 | −.592058 | .340960 | .018867 |
| EXPOSE | .705119 | .050807 | .391191 | .081146 |
| MOY_TP | .827427 | −.411305 | .201033 | .172184 |
| EX1 | −.071405 | .771716 | .287602 | −.182142 |
| EX2 | .599014 | .268014 | .539034 | −.256947 |
| EX3 | .573754 | .410886 | −.068342 | .352293 |
| EX4 | .672017 | .162602 | .028190 | .087754 |
| EX5 | .368246 | .009095 | −.491672 | −.111145 |
| EX6 | .097700 | −.265825 | .582176 | .421531 |
| EX7 | .636349 | −.132226 | −.228022 | .248509 |
| MOY_EXAM | .816595 | .428002 | .119387 | .187593 |
| Expl.Var | 5.933334 | 2.620133 | 1.989705 | 1.799377 |
| Prp.Totl | .312281 | .137902 | .104721 | .094704 |

**fig III.7.1**. Corrélations entre variables et facteurs

Les données du tableau III.7.1 peuvent être représentées de manière graphique. Les figures III.7.2 et III.7.3 représentent ainsi les positions des variables par rapport aux deux pre-

miers facteurs. Ces graphiques sont délivrés respectivement par STATISTICA et **R** (voir chapitre III.5).

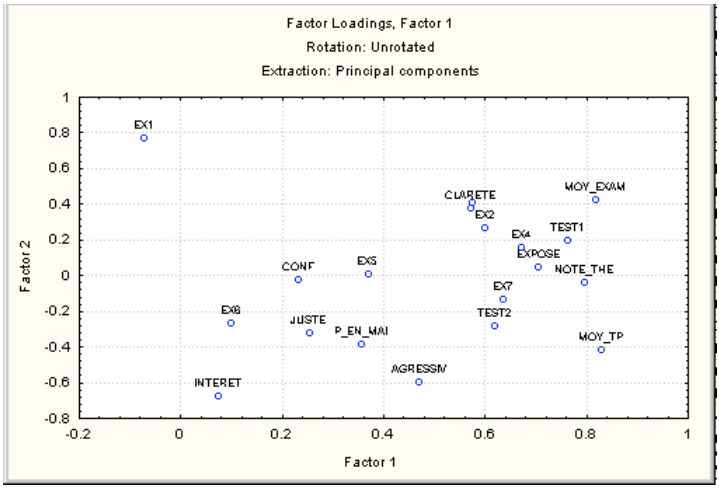



**fig III.7.2**. Plan des facteurs 1 et 2 (STATISTICA)

Les variables à droite sont fortement corrélées positivement au premier facteur. Elles jouent un rôle important dans l'appréciation intuitive de la signification des facteurs. Cette attribution d'une signification peut aussi s'appuyer sur l'opposition entre des variables corrélées positivement et des variables corrélées négativement.

Le premier facteur représente le niveau scolaire de l'étudiant par rapport à la moyenne de la classe (idée de régularité). Cette interprétation est due au fait que le premier facteur est toujours un facteur général. Dans le cas particulier, on le constate par le haut degré de saturation des trois moyennes (MOY_EXAM, NOTE_THE, MOY_TP). On note que la corrélation de EX1 avec ce premier facteur est très faible. Ce qui signifie que la réussite à EX1 est indépendante du niveau général de l'élève.

| |
|---|
| EX1: Quelles sont les trois principales fonctions attendues d'un système cryptographique ? Comment ces fonctions sont-elles réalisées avec un système à clé publique (clé révélée) ? |

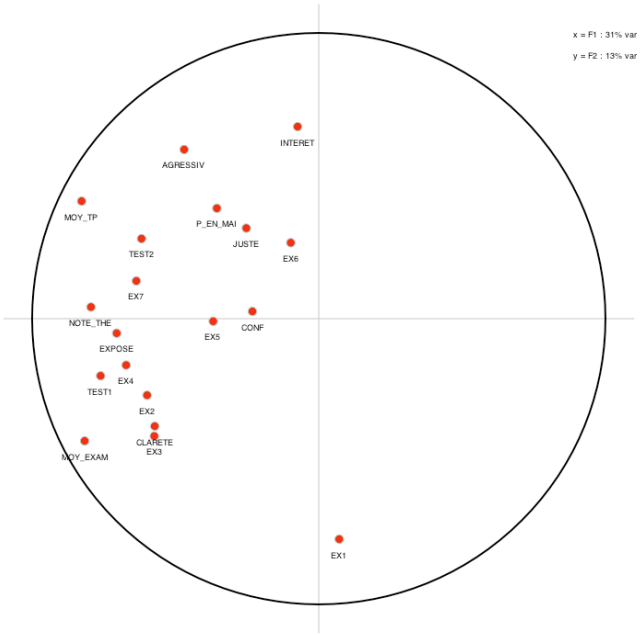



**fig III.7.3**. Diagramme obtenu sous R (commande mdspca de la librairie psy)

Le deuxième facteur oppose EX1 et INTERET. EX1 est une question faisant appel à une référence très précise des notes de cours. INTERET est plutôt lié à un aspect non scolaire (aussi indépendant du niveau général). Le deuxième facteur peut donc être interprété comme l'aspect utilisation de la documentation. Cette interprétation contient aussi l'idée d'un travail important mais peu de recul dans la matière. C'est une connaissance un peu superficielle. La situation respective sur cet axe des trois variables MOY_TP, NOTE_THE et MOY_EXAM colle bien avec cette interprétation.

A noter que seuls ces deux premiers facteurs possèdent un degré de saturation jugé « significatif » par rapport à des variables initiales (devant être supérieur à 0.7 dans ce cas).

Toutefois, il est possible de continuer l'interprétation pour les deux autres facteurs.

Le troisième facteur peut être interprété avec l'idée de « compréhension ». Il correspond par exemple au degré de

réussite à des questions nécessitant un certain degré de synthèse parmi les sujets étudiés, par exemple EX2 et EX6.

| |
|---|
| EX2: Expliquez brièvement l'utilité du protocole MIME. Quel en a été le premier usage ? Donnez quelques exemples liés à l'encodage ("shiftage") des documents. Où retrouve-t-on des informations liées à MIME dans des documents HTML ? |

| |
|---|
| EX6: Une entreprise réalise de la vente par correspondance par Internet. Une « interface » (logée dans un navigateur standard) doit donc être réalisée qui permet de sélectionner des produits, vérifier l'état de la commande, introduire les coordonnées du client, vérifier une certaine cohérence de ces coordonnées, envoyer la commande.<br>Deux solutions extrêmes sont possibles en ce qui concerne l'étape qui précède l'enregistrement de la commande (qui se fera toujours sur le serveur !) : une solution « tout client » (toutes les données sont transmises lors de la première requête sous la forme, par exemple, d'un tableau d'enregistrements), ou une solution « tout serveur ».<br>Donnez quelques avantages et inconvénients de chacune de ces méthodes. Donnez un critère permettant de choisir la solution la mieux adaptée et donnez la description d'une méthode intermédiaire. |

Le quatrième facteur semble lié à la réussite à des items dont l'énoncé représente une certaine difficulté (leur réussite peut donc être liée à une capacité entre compréhension et intuition). EX6, par exemple, est à nouveau parmi les éléments les plus saturés, alors que les tests (notamment QCM) sont à l'opposé. Il peut également correspondre à une « faculté » entre intuition et compréhension.

On peut examiner les autres items en regard de cette interprétation. On voit que, par exemple, EX5 est essentiellement lié au niveau général. Il est neutre par rapport à la bonne utilisation de la documentation, mais que c'est un sujet difficilement traitable par la seule compréhension (en effet, il s'agissait de se souvenir de discussions menée à ce propos

durant les cours). Par ailleurs, l'énoncé ne pose pas de difficulté majeure.

EX5: Des documents sont codés dans un format propriétaire (i.e. qui vous appartient). Vous publiez ces documents sur Internet. Pour les visualiser vos correspondants devront installer des modules logiciels sur leur ordinateur "client". Indiquez des solutions possibles et leurs caractéristiques.

Données : notes-fact.sta 4v * 27c

| VALEURS NUMERIQUES | Résult. Fact. (2ig9899-a | | | |
|---|---|---|---|---|
| | 1 FACTE | 2 FACTEU | 3 FACTEU | 4 FACTEU |
| ROA | 3.5 | 6.0 | 4.9 | 5.3 |
| MIG | 4.5 | 4.1 | 3.3 | 5.0 |
| JVC | 3.6 | 3.8 | 5.8 | 2.8 |
| RDM | 4.3 | 2.7 | 3.4 | 5.1 |
| JOF | 4.5 | 5.9 | 3.6 | 3.1 |
| SOG | 5.2 | 3.9 | 3.7 | 3.7 |
| FLG | 4.5 | 3.3 | 3.9 | 3.7 |
| MOA | 4.8 | 3.9 | 4.8 | 3.8 |
| NEM | 3.5 | 3.5 | 5.2 | 3.7 |
| EMO | 4.7 | 3.5 | 5.3 | 5.2 |
| CHP | 3.4 | 5.4 | 3.8 | 2.7 |
| OLP | 4.7 | 4.5 | 3.5 | 3.3 |
| VIR | 4.8 | 3.4 | 4.0 | 4.4 |
| ANS | 4.1 | 4.4 | 5.1 | 4.3 |
| PHS | 4.3 | 4.4 | 4.0 | 4.3 |
| SEV | 5.1 | 3.1 | 3.5 | 3.3 |
| FVI | 4.6 | 3.6 | 2.5 | 3.7 |
| QUB | 5.7 | 4.2 | 4.5 | 4.0 |
| FAC | 2.9 | 3.8 | 3.9 | 3.2 |
| NDC | 4.4 | 5.1 | 4.5 | 3.8 |
| SDZ | 3.3 | 3.3 | 3.6 | 3.6 |
| CEF | 2.5 | 4.2 | 4.3 | 5.0 |
| SEK | 2.8 | 3.9 | 2.4 | 4.4 |
| DAS | 2.4 | 3.6 | 3.7 | 2.6 |
| MOS | 3.5 | 5.0 | 2.6 | 5.8 |
| JOW | 3.8 | 3.5 | 3.3 | 3.5 |
| SEZ | 2.8 | 2.5 | 5.0 | 5.0 |

**fig III.7.4**. Les notes en facteurs

Finalement pour chacun des 27 étudiants, des notes ont été calculées à partir des notes en facteurs (figure III.7.4). La transformation utilisée est donnée par : y = 0.87 x + 4. Cette normalisation est arbitraire et sert simplement à mettre les notes dans un format familier (entre 1 et 6). On a donc un ensemble de profils d'élèves donnés par un niveau général (régularité), l'utilisation de ressources documentaires (un peu superficielle), l'apport lié à la compréhension et finalement la facilité à traiter des énoncés « complexes ».

- L'observation ROA montre un étudiant relativement peu régulier, un travailleur de la dernière minute, mais qui fait montre d'une compréhension en profondeur.

- L'observation SOG est lié à un cas que l'on pourrait qualifié de scolaire.

- L'observation CHP représente un profil d'un étudiant travailleur mais ayant des difficultés.

- L'observation QUB est maximale du point de `vue` du facteur général.

**Les autres informations délivrées**

Le système informatique délivre de nombreuses informations qui ont été omises précédemment.

*Les valeurs propres*

Les valeurs propres (tableau III.7.1) sont accompagnées de la proportion de variance qu'elles expliquent et des cumuls correspondants (propriété III.7.1).

```
STAT.    Val. Propres (2ig9899-ano.sta)
ANALYSE  Extraction: ACP
FACTOR.
                  % Total   Cumul     Cumul
Valeur  ValPropr  Variance  ValPropr  %age
  1     5.933334  31.22808   5.93333  31.22808
  2     2.620133  13.79018   8.55347  45.01825
  3     1.989705  10.47213  10.54317  55.49038
  4     1.799377   9.47041  12.34255  64.96079
```

**tableau III.7.1**. Valeurs propres

*Les communalités*

Dans le langage de l'analyse factorielle, la proportion de variance d'un item particulier qui est due aux facteurs retenus (qui sont communs ou partagés avec d'autres items) est appelée communalité (ou communauté, en anglais communality). Ces valeurs font l'objet du tableau III.7.2.

```
STAT.    Communautes (2ig9899-ano.sta)
ANALYSE  Extraction: ACP
FACTOR.  Rotation: sans rot.

         Pour 1   Pour 2  Pour 3  Pour 4    R2
Variable Facteur  Fact.   Fact.   Fact.  Multiple

TEST1    .57902   .61906  .64394  .75922   .99783
TEST2    .38238   .46078  .53805  .75848   .99753
NOTE_THE .63152   .63317  .70081  .91765   .99923
CONF     .05355   .05419  .29715  .40926   .85613
JUSTE    .06384   .16427  .18470  .54864   .93746
INTERET  .00543   .45711  .47190  .56697   .89884
CLARETE  .32730   .46894  .58174  .69291   .92879
P_EN_MAI .12636   .27569  .36926  .37920   .88388
AGRESSIV .22047   .57101  .68726  .68762   .95704
EXPOSE   .49719   .49977  .65281  .65939   .97845
MOY_TP   .68464   .85381  .89422  .92387   .99396
EX1      .00510   .60065  .68336  .71654   .84847
EX2      .35882   .43065  .72121  .78723   .85611
EX3      .32919   .49802  .50269  .62680   .85460
EX4      .45161   .47805  .47884  .48654   .70333
EX5      .13561   .13569  .37743  .38978   .74541
EX6      .00955   .08021  .41914  .59683   .84135
EX7      .40494   .42242  .47442  .53618   .80020
MOY_EXAM .66683   .85001  .86427  .89946   .97172
```

**tableau III.7.2**. Communautés

On constate que MOY_TP est la variable la mieux représentée par les quatre facteurs, et P_EN_MAI la moins bien représentée.

*Composants des facteurs*

Les facteurs peuvent s'exprimer à l'aide des variables et vice-versa sous forme de combinaisons linéaires. La première colonne du tableau III.7.3 donne les composantes du premier facteur selon les variables primitives réduites. Chaque colonne du tableau III.7.3 est proportionnelle à la co-

lonne correspondante du tableau de la figure III.7.1, le facteur de proportionnalité étant la valeur propre correspondante (propriété III.7.2).

```
STAT.    Coeff. des resultats factoriels
(2ig9899-ano.sta)
ANALYSE  Rotation: sans rot.
FACTOR.  Extraction: ACP

            Facteur    Facteur    Facteur     Facteur
Variable      1          2          3          4

TEST1       .128248    .076365   -.079276   -.188693
TEST2       .104219   -.106862   -.139711   -.260921
NOTE_THE    .133935   -.015510   -.130709   -.258794
CONF        .038999   -.009658    .247733   -.186083
JUSTE       .042584   -.120950   -.071840    .335268
INTERET     .012422   -.256501    .061123   -.171356
CLARETE     .096422    .143636   -.168797    .185299
P_EN_MAI    .059911   -.147486   -.153737    .055416
AGRESSIV    .079137   -.225965    .171362    .010486
EXPOSE      .118840    .019391    .196607    .045097
MOY_TP      .139454   -.156979    .101037    .095691
EX1        -.012035    .294533    .144545   -.101225
EX2         .100957    .102290    .270912   -.142798
EX3         .096700    .156819   -.034348    .195786
EX4         .113261    .062059    .014168    .048769
EX5         .062064    .003471   -.247108   -.061768
EX6         .016466   -.101455    .292594    .234265
EX7         .107250   -.050466   -.114601    .138108
MOY_EXAM    .137628    .163351    .060002    .104255
```

**tableau III.7.3**. Composantes des facteurs

*Les notes en facteurs*

En utilisant les coefficients précédents et les variables (normalisées) il est possible de calculer les notes en facteur (figure III.7.5). La moyenne de chacune de ces notes vaut 0 et l'écart-type 1.

**Conclusion**

En définitive, on voit que des outils principalement utilisés pour la "recherche", peuvent s'avérer utiles dans l'enseignement de tous les jours. De nombreux logiciels existent qui permettent de faire aisément ces analyses, mais la partie délicate reste l'interprétation qui devrait toujours rester nuancée.

**Pour réfléchir**

1) Trouver des interprétations pour les quatre autres facteurs

2) Trouver et interpréter des profils typiques d'étudiants.

Résult. Fact. (2ig9899-ano.sta)

ANALYSE FACTOR.
Rotation: sans rot.
Extraction: ACP

| Obs. | Facteur 1 | Facteur 2 | Facteur 3 | Facteur 4 |
|---|---|---|---|---|
| ROA | -.58972 | 2.30123 | 1.05415 | 1.53563 |
| MIG | .52053 | .15169 | -.76062 | 1.14957 |
| JVC | -.52163 | -.27228 | 2.11267 | -1.43379 |
| RDM | .38948 | -1.54028 | -.65808 | 1.26438 |
| JOF | .52064 | 2.19465 | -.47699 | -1.00408 |
| SOG | 1.37203 | -.15602 | -.37175 | -.38845 |
| FLG | .56487 | -.85597 | -.16511 | -.37406 |
| MOA | .87893 | -.18334 | .88095 | -.22281 |
| NEM | -.57393 | -.59882 | 1.39223 | -.34349 |
| EMO | .84198 | -.54991 | 1.43920 | 1.41776 |
| CHP | -.74151 | 1.55450 | -.21085 | -1.54886 |
| OLP | .77068 | .51306 | -.58029 | -.79592 |
| VIR | .94544 | -.72121 | .04412 | .45266 |
| ANS | .11016 | .44563 | 1.30930 | .28477 |
| PHS | .29730 | .42084 | -.01948 | .28586 |
| SEV | 1.26132 | -1.04700 | -.54195 | -.77907 |
| FVI | .67885 | -.44045 | -1.69939 | -.40153 |
| QUB | 1.93791 | .21112 | .51061 | .02598 |
| FAC | -1.27385 | -.23265 | -.15529 | -.89086 |
| NDC | .44073 | 1.24229 | .56072 | -.23925 |
| SDZ | -.77747 | -.83663 | -.43667 | -.52172 |
| CEF | -1.72598 | .26750 | .28940 | 1.10961 |
| SEK | -1.37642 | -.17707 | -1.85239 | .48786 |
| DAS | -1.79818 | -.52333 | -.33887 | -1.61587 |
| MOS | -.54463 | 1.10019 | -1.58983 | 2.03585 |
| JOW | -.25082 | -.58595 | -.86123 | -.62322 |
| SEZ | -1.35670 | -1.68178 | 1.12542 | 1.13302 |

**fig III.7.5**. Les notes en facteurs (variables réduites)

## Chapitre III.8. Exemple d'analyse factorielle des correspondances (AFC)

### Contexte

Une étude de la presse a été menée pour connaître les représentations de l'usage de l'Internet à l'école véhiculées par la presse[1]. Pour l'analyse de contenu, un peu plus de cent articles (103) ont été retenus, extraits de divers types de presse (quotidiens, hebdomadaires, presse pédagogique, etc.). A la lecture de ces articles, 244 descripteurs ont été sélectionnés pour indexer le corpus. L'étude voulait aussi explorer quelques outils applicables à ce genre d'étude. Deux analyses hiérarchiques ont été réalisées, l'une sur les articles (utilisant les descripteurs comme index) et l'autre sur les descripteurs (en utilisant les articles comme index). Cela a permis de dégager 10 groupes d'articles et 17 concepts (regroupements de descripteurs).

### L'étude

Il est honnête de faire l'hypothèse que ces deux catégorisations ne sont pas indépendantes et l'idée ici est d'estimer ce degré de dépendance.

Pour cela on établit une table de contingence (tableau III.8.1) dans lequel les groupes figurent en lignes et les concepts en colonnes. Au croisement de la ligne i et de la colonne j se trouve le nombre de descripteurs du concept j décrivant des articles du groupe i.

Le *chi-2* total vaut 1227.24 (ddl=144, $p$=0.000). Il montre que la répartition des effectifs est loin d'être le fruit du hasard. Comment mieux caractériser cette dépendance entre groupes et concepts ?

L'analyse factorielle des correspondances offre une réponse à cette question. Elle va mettre en évidence des profils de groupes (en fonction des composantes représentées par le nombre de concepts) et, de façon duale, des profils de concepts. Le mode de composition des profils est additif[2].

---

[1] Données reprises de Berney & Pochon (2000).

[2] Les analyses « log-linéaires » (chapitre III.9) privilégient un mode multiplicatif.

| | CO1 | CO2 | CO3 | CO4 | CO5 | CO6 | CO7 | CO8 | CO9 |
|---|---|---|---|---|---|---|---|---|---|
| g1 | 31 | 9 | 17 | 23 | 26 | 24 | 44 | 7 | 6 |
| g2 | 16 | 6 | 30 | 40 | 20 | 17 | 26 | 14 | 39 |
| g3 | 20 | 14 | 62 | 96 | 41 | 15 | 30 | 14 | 27 |
| g3' | 34 | 19 | 25 | 46 | 22 | 15 | 17 | 14 | 8 |
| g4 | 14 | 3 | 7 | 11 | 8 | 21 | 32 | 48 | 13 |
| g5 | 47 | 30 | 24 | 15 | 14 | 13 | 62 | 37 | 7 |
| g6 | 37 | 62 | 15 | 25 | 22 | 12 | 18 | 23 | 1 |
| g7 | 33 | 72 | 3 | 9 | 12 | 5 | 13 | 10 | 1 |
| g8 | 16 | 18 | 24 | 20 | 12 | 4 | 17 | 10 | 5 |
| g9 | 21 | 19 | 12 | 18 | 17 | 15 | 10 | 14 | 6 |
| total | 269 | 252 | 219 | 303 | 194 | 141 | 269 | 191 | 113 |

| | CO10 | CO11 | CO12 | CO13 | CO14 | CO15 | CO16 | CO17 |
|---|---|---|---|---|---|---|---|---|
| g1 | 33 | 9 | 21 | 8 | 10 | 8 | 4 | 4 |
| g2 | 11 | 5 | 18 | 13 | 5 | 3 | 7 | 8 |
| g3 | 18 | 6 | 36 | 8 | 21 | 17 | 10 | 5 |
| g3' | 6 | 10 | 31 | 6 | 20 | 6 | 17 | 11 |
| g4 | 4 | 0 | 13 | 7 | 4 | 1 | 3 | 6 |
| g5 | 8 | 4 | 19 | 25 | 15 | 16 | 17 | 27 |
| g6 | 8 | 6 | 25 | 5 | 6 | 25 | 60 | 14 |
| g7 | 2 | 3 | 21 | 1 | 5 | 14 | 46 | 31 |
| g8 | 7 | 3 | 32 | 11 | 26 | 24 | 9 | 14 |
| g9 | 1 | 0 | 23 | 1 | 1 | 20 | 15 | 1 |
| total | 98 | 46 | 239 | 85 | 113 | 134 | 188 | 121 |

**tableau III.8.1**. Table de contingence donnant le nombre de concepts présents dans chaque groupe d'articles

### Les résultats

Les résultats délivrés par STATISTICA sont les suivants. Tout d'abord les valeurs propres (tableau III.8.2) qui permettent de calculer le pourcentage de la variation (inertie) « expliquée » par les axes. Ici les deux premières dimensions expliquent 67% de l'inertie totale. La liaison entre groupes et concepts est donc essentiellement dépendante de deux paramètres.

D'autres informations, les contributions de chaque groupe et de chaque concept à l'inertie ne sont pas reproduites ici.

Les données suivantes sont les coordonnées des groupes et des concepts dans l'axe des facteurs. Le tableau III.8.3 présente les coordonnées des groupes. Ces coordonnées permettent de représenter groupes et concepts dans un système d'axes. La figure III.8.1 est produite par STATISTICA. La

figure III.8.2 est obtenue sous **R** avec la « librairie » FactoMineR (voir les activités pour réfléchir).

```
Analyse des Correspondances d'une Table d'ordre
deux

Nombre de variables (colonnes de la table): 17
Nombre d'obs. actives (lignes de la table): 10

Valeurs propres: .1867 .0884 .0444 .0354 .0263
.0129 .0081 .0057 .0045
Chi-deux Total=1227.24 ddl=144 p=0.000

STAT.    Valeurs Propres et Inertie de toutes les
Dimensions (jesa.sta)
ANALYSE  Table d'Entree (Lignes x Colonnes): 10 x
17
CORRESP. Inertie Totale=.41252 Chi2=1227.2
ddl=144 p=0.0000

Nombre   Valeurs  ValPropr %age    %age     Chi-2
de Dims. Singul.  Singul.  Inertie  Cumule

  1      .43209* .18670* 45.2591* 45.259* 555.44*
  2      .29731* .08840* 21.4282* 66.687* 262.98*
  3      .21083  .04445  10.7751  77.462  132.24
  4      .18819  .03541   8.5847  86.047  105.36
  5      .16223  .02632   6.3796  92.427   78.29
  6      .11351  .01289   3.1234  95.550   38.33
  7      .09012  .00812   1.9689  97.519   24.16
  8      .07578  .00574   1.3920  98.911   17.08
  9      .06703  .00449   1.0890 100.000   13.36
```

**tableau III.8.2**. Les valeurs propres

Le pas suivant consiste à donner une interprétation aux facteurs. Cela nécessite de bien connaître les caractéristiques des objets (groupes et concepts) en présence. Avec toutes les précautions d'usage, ici le premier axe oppose des cas pratiques (par exemple l'organisation des net@days) à des réflexions plus théoriques. Le deuxième axe oppose les problèmes de sociétés à ceux plus spécifiques du monde scolaire. La proximité géométrique de deux éléments (par exemple g4 et CO8) peut correspondance à une proximité au niveau des significations.

```
STAT.    Coord. Lignes et Contributions a l'Iner-
tie
ANALYSE  Table d'Entree (Lignes x Colonnes): 10 x
17
CORRESP. Standardisation: Profils lignes et co-
lonnes

Nom   Ligne Coord.  Coord.      Inertie
Ligne Num.  Dim.1   Dim.2 Masse Qualite Relative

g1       1  -.31854  .07867 .09546 .23844 .10449
g2       2  -.51081 -.02454 .09345 .57490 .10305
g3       3  -.43851 -.36648 .14790 .91202 .12839
g3b      4  -.08567 -.18222 .10319 .32306 .03139
g4       5  -.31617  .80219 .06555 .83578 .14134
g5       6   .07379  .38197 .12773 .60981 .07685
g6       7   .55565 -.08175 .12235 .85181 .10983
g7       8   .90361 -.05658 .09445 .92811 .20223
g8       9   .00768 -.16122 .08471 .08894 .06014
g9      10   .14032 -.07019 .06521 .09202 .04229
```

**tableau III.8.3**. Les coordonnées des groupes dans le système des facteurs

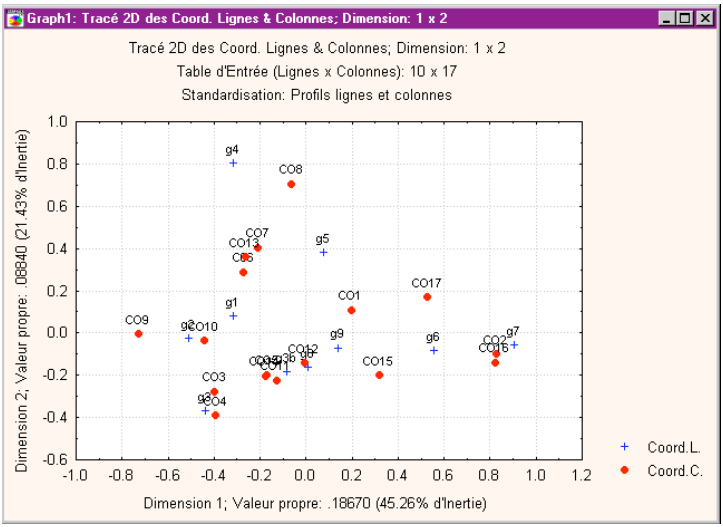



**fig III.8.1**. Représentation des groupes et des concepts dans le même système d'axe (STATISTICA)

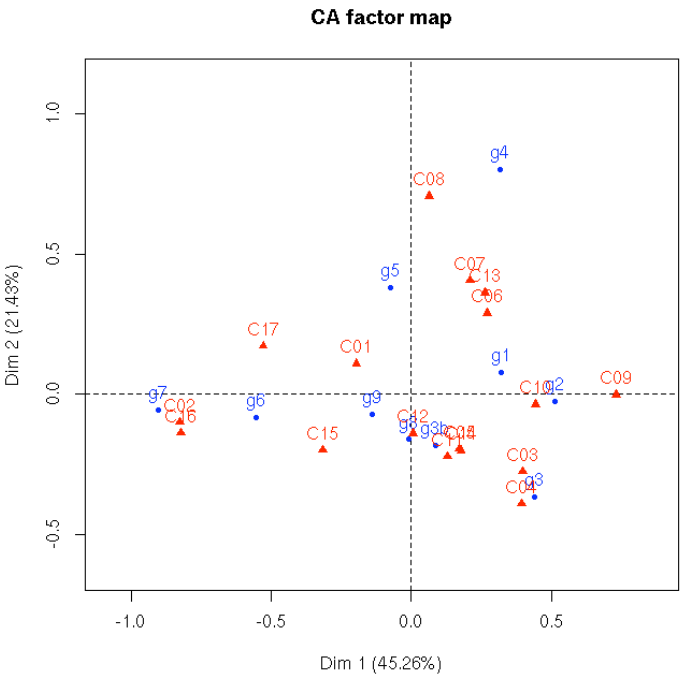



**fig III.8.2**. Représentation des groupes et des concepts dans le même système d'axe (**R** avec FactoMineR)

## Pour réfléchir

*Réaliser quelques analyses factorielles*

Pour réaliser des analyses factorielles (ACP, AFC, ACM, etc.) le package FactoMineR[1] est souvent considéré comme un standard. Les manipulations de base sont le chargement du package (à réaliser une fois pour toutes) et l'activation de la librairie.

```
> install.packages("FactoMineR")
> library(FactoMineR)
```

Pour entrer en matière, reproduisez la figure III.8.2 en chargeant les données du tableau III.8.1 et en lançant l'analyse.

```
> load("tabAFC.dat")
> CA(df.article)
```

[1] http://factominer.free.fr (consulté : septembre 2012).

Puis explorez les paramètres de la commande `CA` et les autres analyses possibles.

```
> help(CA)
```

*Acquérir de l'intuition en analyse hiérarchique*

Cette activité se réfère à l'analyse hiérarchique présentée dans le chapitre I.4 qui a servi à produire les données exploitées dans ce chapitre I.4.

Les données présentées sont disponibles pour **R** dans le fichier auteurDoc.txt[1].

Effectuez diverses analyses sur ces données en variant les distances et les méthodes d'agrégation et comparer les résultats. Par exemple :

```
> dist_eucl <- dist(auteurDoc,method="euclidean")
> clust_eucl_w <- hclust(dist_eucl,method="ward")
> plot(clust_eucl_w,main="Analyse: auteurDoc-1")
```

Le même exercice peut s'effectuer sur vos propres données.

[1] Commande : `auteurDoc <- read.table("auteurDoc.txt")`

## Chapitre III.9. Analyse log-linéaire[1]

L'analyse log-linéaire est en quelque sorte une version multiplicative de l'analyse factorielle des correspondances ou de l'analyse de variance. Elle s'applique à des tables de contingence de dimension 2 ou plus et permet de caractériser les relations entre des caractères nominaux (variables qualitatives) selon un modèle d'association (et non de régression, il n'y a pas de distinction entre variables explicatives et variables à expliquer).

Dans le cas général, le modèle traite $k$ caractères $C_i$, chacun avec $I_{Ci}$ modalités et des effectifs $n_{i1,i2,\ldots,ik}$ (nombre d'individus prenant la modalité $i_j$ pour le caractère $C_i$).

Pour simplifier la présentation, on se limitera à deux caractères A et B, de modalités respectives ($a_1$,…, $a_n$) et ($b_1$,…, $b_n$). Les effectifs $n_{ij}$ constituent une table de contingences classique (tableau III.9.1).

| | $b_1$ | $b_2$ | … | $b_m$ | Total |
|---|---|---|---|---|---|
| $a_1$ | $n_{11}$ | $n_{21}$ | … | $n_{m1}$ | $n_{.1}$ |
| … | … | … | … | … | … |
| $a_n$ | $n_{1n}$ | $n_{2y}$ | … | $n_{mn}$ | $n_{.n}$ |
| Total | $n_{1.}$ | $n_{2.}$ | … | $n_{m.}$ | $N$ |

**tableau III.9.1**. Table de contingence liés aux caractères A et B

Le but de l'analyse est, comme dans le cas des analyses factorielles, de décrire ces données avec parcimonie sur la base de l'imposition d'une structure particulière. L'analyse log-linéaire impose une structure multiplicative, par exemple de trouver des paramètres $\sigma_j$ tels que des coefficients $\nu_{ij} = \nu\sigma_j$, qui ne dépendent que de la ligne, approchent $n_{ij}$ « au mieux » selon la statistique $G^2$ (voir chapitre III.4).

On se ramène à un modèle linaire en passant par le logarithme (d'où le nom de l'analyse) :

$$\nu_{ij} = \nu\sigma_i \Leftrightarrow \log(\nu_{ij}) = \log\nu + \log\sigma_i = \mu + \alpha_i \text{ avec la contrainte : } \sum\alpha_i = 0$$

[1] Dans cette annexe, la notation log représente le logarithme naturel.

Le tableau III.9.2 donne tous modèles possibles dans le cas de deux caractères avec $K = nm$ (nombre de cellules).

| Les modèles | | ddl | Remarques |
|---|---|---|---|
| Multiplicatifs | Additifs | | |
| 1) $\nu_{ij} = \nu$ | $\log(\nu_{ij}) = \mu$ | $K$-1 | Les valeurs des cellules sont identiques (effet général). |
| 2) $\nu_{ij} = \nu\sigma_i$ | $\log(\nu_{ij}) = \mu + \alpha_i$ | $K$-$n$ | Les valeurs d'une même ligne sont toutes identiques (effet ligne). |
| 3) $\nu_{ij} = \nu\tau_j$ | $\log(\nu_{ij}) = \mu + \beta_j$ | $K$-$m$ | Les valeurs d'une même colonne sont identiques (effet colonne). |
| 4) $\nu_{ij} = \nu\sigma_i\tau_j$ | $\log(\nu_{ij}) = \mu + \alpha_i + \beta_j$ | $(n$-1$)(m$-1$)=$ $K$-$m$-$n$ +1 | Modèle d'indépendance . |
| 5) $\nu_{ij} = \nu\sigma_i\tau_j\upsilon_{ij}$ | $\log(\nu_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$ | 0 | Modèle complet (saturé). |

**tableau III.9.2**. Les différents modèles d'une analyse log-linéaire

Dans ce tableau, $\mu$ représente l'effet général, $\alpha_i$ l'effet ligne, $\beta_j$ l'effet colonne et $\gamma_{ij}$ l'effet d'interaction avec les conditions :

$$\sum \alpha_i = \sum \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

Le choix du modèle et le calcul des coefficients sont effectués de façon à minimiser la valeur de $G^2$. Dans le cas à deux caractères, il suffit de résoudre un système d'équations linéaires pour chaque cas. Dans le cas général, des algorithmes itératifs permettent de faire le travail.

De façon plus précise, les différents cas seront chaque fois commentés et illustrés par un exemple inspiré de Kah & Pruvot (2003). Dans cet exemple fictif, 285 personnes sont interrogées pour savoir si elles sont favorables à ce que l'école héberge les enfants pour la pause de midi. Les répon-

ses sont réparties selon la provenance des personnes de zones rurale ou urbaine. Le tableau III.9.3 résume les données.

| | Rural | Urbain | Total |
|---|---|---|---|
| Oui | 65 | 65 | 130 |
| Non | 107 | 48 | 155 |
| Total | 172 | 113 | 285 |

**tableau III.9.3**. Données à analyser

*Cas 1 :* $\log(v_{ij}) = \mu$

C'est le cas de l'équiprobabilité entre les cellules. Le nombre de degrés de liberté est $K$-1, il n'y a qu'une contrainte pour le choix des valeurs des $K$ cellules : le nombre total d'observations. L'effectif théorique, le même dans chaque cellule, est nécessairement $N / K$.

| | Rural | Urbain | Total |
|---|---|---|---|
| Oui | 71.25 | 71.25 | 142.5 |
| Non | 71.25 | 71.25 | 142.5 |
| Total | 142.5 | 142.5 | 285 |

**tableau III.9.4**. Les effectifs théoriques selon le premier modèle

Dans ce cas (tableau III.9.4) : $\mu = \log(71.25)$ ; $G^2 = 25.230$ ; ddl = 3 ; $p = 0.00001$. La faible valeur de $p$ incite à rejeter le modèle. On notera donc l'emploi particulier du test statistique. On attend une p-value non significative pour retenir une solution.

*Cas 2 :* $\log(v_{ij}) = \mu + \alpha_i$

C'est le cas de l'équiprobabilité entre les modalités Rural/Urbain. Chaque ligne est indépendante. Le nombre de degrés de liberté vaut donc $N$-$n$ (une contrainte par ligne). La somme des $\alpha_i$ doit être nulle. C'est la contrainte sur les paramètres. Comme les valeurs des lignes sont constantes, toutes les sommes des colonnes sont également égales entre elles.

Les valeurs « théoriques » s'obtiennent en résolvant le système :

$$\begin{cases} \alpha_1 + \alpha_2 = 0 \\ \mu + \alpha_1 = \log\dfrac{130}{2} \\ 4\mu + 2(\alpha_1 + \alpha_2) = 4\mu = \log(65 \times 65 \times 77.5 \times 77.5) \end{cases}$$

| | Rural | Urbain | Total |
|---|---|---|---|
| Oui | 65 | 65 | 130 |
| Non | 77.5 | 77.5 | 155 |
| Total | 142.5 | 142.5 | 285 |

**tableau III.9.5**. Les effectifs théoriques selon le deuxième modèle

Dans ce cas (tableau III.9.5) : $\mu$ = log 70.975 = 4.26 ; $\alpha_1 = -\alpha_2$ = log 0.916 = -0.088; $G^2$ = 23.035 ; ddl = 2 ; $p$ = 0.00001. La faible valeur de $p$ incite à rejeter le modèle.

*Cas 3 :* $\log(v_{ij}) = \mu + \beta_j$

C'est le cas de l'équiprobabilité entre les modalités Oui/Non. Chaque colonne est indépendante. Le nombre de degrés de liberté vaut donc $N$-$m$ (une contrainte par colonne). La contrainte sur les paramètres est que la somme des $\beta_j$ doit être nulle. Comme les valeurs des colonnes sont constantes, toutes les sommes des lignes sont également égales entre elles.

| | Rural | Urbain | Total |
|---|---|---|---|
| Oui | 86 | 56.5 | 142.5 |
| Non | 86 | 56.5 | 142.5 |
| Total | 172 | 113 | 285 |

**tableau III.9.6**. Les effectifs théoriques selon le troisième modèle

Dans ce cas (tableau III.9.6) : $\mu$ = log(69.707) = 4.24 ; $\beta_1 = -\beta_2$ = log 1.234 = 0.210 ; $G^2$ = 12.928 ; ddl = 2 ; $p$ = 0.00156. La faible valeur de $p$ incite à rejeter le modèle.

A noter que ce modèle est plus vraisemblable que le précédent puisque la répartition Rural/Urbain est ici fixe alors que l'opinion Oui/Non relève du domaine aléatoire.

*Cas 4 :* $\log(\nu_{ij}) = \mu + \alpha_i + \beta_j$

Ce modèle avec la somme des $\alpha_i$ et la somme des $\beta_j$ nulles implique l'indépendance des caractères. Les totaux marginaux sont fixés, ce qui représente *n*+*m* contraintes. Le nombre de degrés de liberté vaut donc *N*-*n*-*m*.

| | Rural | Urbain | Total |
|---|---|---|---|
| Oui | 78.46 | 51.54 | 130 |
| Non | 93.54 | 61.46 | 155 |
| Total | 172 | 113 | 285 |

**tableau III.9.7**. Les effectifs théoriques selon le modèle d'indépendance

Dans ce cas (tableau III.9.7) : $\mu = \log(69.438) = 4.24$ ; $\alpha_1 = -\alpha_2 = \log 0.916 = -0.088$ ; $\beta_1 = -\beta_2 = \log 1.234 = 0.210$ ; $G^2 = 10.732$ ; ddl = 1 ; $p = 0.00105$. La faible valeur de *p* incite encore à rejeter le modèle.

Seul le modèle saturé conviendra qui rend compte exactement des valeurs observées. Les données n'ont pas pu être réduites. Par contre, on pourra examiner les relations entre les caractères.

*Cas 5 :* $\log(\nu_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$

Ce modèle, avec la somme des $\alpha_i$, la somme des $\beta_j$ et la somme des $\gamma_{ij}$ (sur les indices *i* et sur les indices *j*) nulles, permet d'obtenir les coefficients du tableau III.9.3. C'est le modèle saturé. Il n'y a plus de degré de liberté. $G^2 = 0$. Dans ce cas $\mu = (1/4) \log(65 \times 65 \times 107 \times 48) = \log 68.252 = 4.223$. Les autres coefficients sont représentés dans le tableau III.9.8. Ils s'obtiennent facilement en résolvant des équations linéaires.

| | Rural | Urbain | Total |
|---|---|---|---|
| Oui | $\gamma_{11} = -0.2004$ | $\gamma_{12} = 0.2004$ | $\alpha_1 = -0.0488$ |
| Non | $\gamma_{21} = 0.2004$ | $\gamma_{22} = -0.2004$ | $\alpha_2 = 0.0488$ |
| Total | $\beta_1 = 0.2004$ | $\beta_2 = -0.2204$ | $\mu = 4.223$ |

**tableau III.9.8**. Les différents coefficients dans le cas du modèle saturé

Dans ce cas, les coefficients $\alpha$ et $\beta$ nous indiquent simplement les différences d'effectifs entre les modalités (faibles entre Oui/Non plus forte entre Rural/Urbain). Le coefficient $\gamma$ donne une indication sur l'interaction. Il indique une défection du milieu rural pour le « oui ».

Ces conclusions mettent en évidence que l'analyse loglinéaire, comme la plupart des analyses se basant sur les effectifs, ne donne pas directement accès aux individus. Ce sont les hypothèses expérimentables qui peuvent être déclinées en terme d'effectifs qui bénéficient des techniques loglinéaires.

## Chapitre III.10. Brève introduction à l'analyse de variance

### Introduction

L'analyse de variance (ANOVA) est une technique statistique classique introduite par R. A. Fisher permettant de tester des différences entre groupes. Plus précisément, si $G_i$ ($i$=1,…,$k$) sont $k$ groupes de, respectivement, $n_i$ individus pour lesquels est défini un caractère à échelle d'intervalle $X$, il s'agit de « tester » si les moyennes $m_i$(X), moyennes de $X$ sur $G_i$, sont différentes.

Dans cette procédure, la variation totale d'un ensemble de données (autour d'une moyenne) est décomposée en une combinaison linéaire de composantes

Cette technique est principalement utilisée dans le cadre du paradigme population/échantillon pour estimer si les observations sur des échantillons peuvent être étendues à la population. Si, dans la technique classique, les groupes sont constitués d'individus au sens propre (des étudiants, par exemple), la théorie peut s'étendre à d'autres populations (ensemble d'items, par exemple). On entre alors dans la théorie de la généralisabilité (Cardinet, Johnson & Pini, 2010).

Dans l'analyse de la variance, les hypothèses d'artefact sont assez nombreuses. Notamment, cela suppose une distribution normale de $X$ et des conditions sur les écarts-type de $X$ à l'intérieur des groupes (voir le complément 1). Toutefois, la théorie est relativement robuste. On peut aussi de ce fait, l'utiliser dans le cadre du paradigme des petits groupes.

### Technique de base

Les briques de base sont simples. Comme son nom l'indique, l'élément de base est la variance.

Si G est un groupe de $n$ individus et un caractère à échelle d'intervalle $X$ prenant les valeurs $x_i$ sur les différents individus $i$ de G on a :

- la moyenne : $m(X) = \frac{1}{n} \sum_{i=1,..n} x_i$ ;

- la somme des carrés des écarts ou plus simplement somme des carrés (*sum of squares*) : $SS(X) = \sum_{i=1,..n} (x_i - m(X)^2)$ ;

- la somme des carrés moyens (ou variance): $MS(X) = \frac{SS(X)}{n-1}$

où $n$-1 est le nombre de degrés de liberté de $X$ (la moyenne est considérée comme fixée).

**Analyse de variance simple**

Soit un groupe G d'individus divisé en $k$ sous-groupes (catégories) $G_i$. Cette division est liée à un caractère nominal appelé facteur, ou parfois facette, dans le cadre de l'analyse de la variance. Le nombre total d'individus est noté $n$ et le nombre d'individus de la catégorie $G_i$ : $n_i$.

On peut alors considérer :

- la moyenne ($m(X)$), la somme des carrés ($SST(X)$), le nombre de degrés de liberté ($n$-1) et le carré moyen $MST(X) = SST(X)/(n\text{-}1)$ pour l'ensemble des individus. Le symbole $SST(X)$ est utilisé pour signifier somme des carrés totale.

- la moyenne, la somme des carrés pour chacune des $k$ catégories : $m_i(X)$, $SS_i(X)$. La somme des $SS_i(X)$ est appelée somme des carrés *dans* les groupes. Elle est notée $SSd(X)$. Le nombre de degrés de liberté correspondant est $n$-$k$ ($k$ moyennes sont fixées).

De plus, on considère la somme des carrés *entre* les groupes calculé de la façon suivante : $SSe(X) = \sum_{i=1,..k} n_i (m_i(X) - m(X))^2$.

Cela revient à attribuer à chaque individu la moyenne de son groupe. Le nombre de degrés de liberté est $k$-1 ($k$ moyennes avec la moyenne des moyennes fixée).

Un calcul technique mais direct (voir le complément 2) montre le résultat fondamental suivant :

$$SST(X) = SSe(X) + SSd(X) \quad \text{[III.10.I]}$$

En clair : la somme totale des carrés se décompose en les carrés entre les groupes et les carrés dans les groupes.

Cette formule exprimée à l'aide des carrés moyens devient :

$$MST(X)(n-1) = MSe(X)(k-1) + MSd(X)(n-k)$$

La variance qui différentie les groupes est *MSe*(*X*). Elle est dite variance de différentiation. *MSd*(*X*) est la variance de généralisation ou terme d'erreur. L'attribution d'une signification à « l'erreur », entre fluctuation de l'instrument et fluctuation propre au caractère mesuré, est une opération délicate souvent passée sous silence (Laurencelle & Allaire, 1998).

Plus le rapport *F*=*Mse* / *MSd* est grand, plus la différentiation est grande et plus le terme d'erreur est petit.

En l'absence de différence systématique entre les groupes (hypothèse nulle), *F* s'approchera de 1 (variance *entre* groupes vaut la variance entre individus). Avec $MSe(X) \approx MSd(X)$ on trouve :

$$MST(X)(n-1) \approx MSd(X)(k-1) + MSd(X)(n-k) = MSd(X)(n-1)$$

$$MST(X) \approx MSd(X)$$

Dans ce cas, la variance *dans* les groupes mesure la variance totale.

La question que l'on peut poser est pourquoi avoir choisi ces coefficients. Il se trouve que moyennant le fait que *X* se distribue normalement, *MSe* et *MSd* suivent une distribution du *chi-2*. Il est alors possible de calculer une distribution théorique du rapport de deux distributions qui respectent la loi du *chi-2*. C'est la loi *F* calculée par Fisher et Snedecor. Cette distribution dépend de deux paramètres, le nombre de degrés de liberté du numérateur et du dénominateur.

**Pratiquement**

Effectuer les calculs revient à faire des sommes de carrés. Quant à loi *F*, elle est programmée dans la plupart des tableurs (dans `EXCEL : LOI.F(valeur,df1,df2)`). Pour **R**, la fonction correspondante est `pf` (il faut calculer `1 - pf(valeur,df1,df2)`).

Mais les logiciels de statistique permettent de réaliser directement une analyse de variance. A titre d'exemple voici une analyse de variance sur le jeu de données du questionnaire (chapitre III.6) qui permet de comparer les opinions (`op2`) sur les trois groupes d'âge dont les moyennes sont données dans le tableau III.10.1.

```
mean(op2[age==1])
[1] 3.166667
mean(op2[age==2])
[1] 2.444444
mean(op2[age==3])
[1] 2.833333
```

**tableau III.10.1**. Moyenne de la variable op2 pour les trois groupes d'âge

De fait, plusieurs possibilités existent dans **R** pour effectuer l'analyse. La plus simple est d'enchâsser les deux commandes `aov` et `summary` . La fonction `aov` procède au calcul des carrés *entre* et *dans* les groupes (noté residuals) de `op2` selon le facteur `age`. Cette propriété doit être dûment précisée afin de permettre de déterminer le nombre de degrés de liberté. La fonction `summary` coordonne les résultats dans un format de présentation standard.

```
> summary(aov(op2 ~ factor(age), quest))
             Df Sum Sq  Mean Sq F value  Pr(>F)
factor(age)   2  4.704   2.352  1.5098   0.2307
Residuals    51 79.444   1.558
```

**listing III.10.1**. Résultat d'une analyse de variance

Du point de vue des notations, l'usage du statisticien est de présenter les résultats en tableau (listing III.10.1). *F* est alors indiqué comme en-tête de colonne. Le mathématicien préférera peut-être l'écriture fonctionnelle $F_{1,48}(1.5098)$ (fonction d'une variable avec deux paramètres). L'informatique adopte l'écriture fonctionnelle à trois variables.

### Analyse de la variance à deux dimensions

Le procédé se généralise aux dimensions supérieures. Le tableau III.10.2 présente les moyennes des groupes obtenus par croisement des caractères « age » et « sexe »[1].

| Age \ sexe | 1 (jeune) | 2 (moyen) | 3 (âgé) | total |
|---|---|---|---|---|
| 1 (F) | 4.11 | 2.11 | 1.67 | 2.63 |
| 2 (M) | 2.22 | 2.78 | 4 | 3 |
| Total | 3.17 | 2.44 | 2.83 | 2.81 |

**tableau III.10.2**. Moyenne de la variable op2 selon le sexe et les trois groupes d'âge

[1] La moyenne de la première case peut se calculer par la commande **R** : `mean(op2[sexe==1 & age==1])`

La formule fondamentale III.10.I est également valable et peut encore se décomposer, dans le cas où les effectifs des groupes sont égaux, en :

$$SST(X) = SSc(X) + SSr(X) + SSi(X) + SSd(X)$$

Si les effectifs des cellules ne sont pas égaux, le résultat de la procédure standard va dépendre de l'ordre dans lequel l'analyse est menée (voir le complément 3). En cas d'inégalité, on peut retirer aléatoirement des individus de chaque groupe. Il existe également des méthodes alternatives, notamment si les effectifs des lignes sont proportionnels. Le listing III.10.2 donne le résultat de l'analyse avec l'effet simple selon sexe et âge et l'effet d'interaction.

```
> summary(aov(op2 ~ factor(sexe)*factor(age)))
             Df Sum Sq Mean Sq F value  Pr(>F)
sexe          1  1.852  1.852   2.4096  0.12716
age           2  4.704  2.352   3.0602  0.05612 .
sexe:age      2 40.704 20.352  26.4819  1.7e-08 ***
Residuals    48 36.889  0.769
```

**listing III.10.2**. Résultat de l'analyse de variance

Si l'âge ne semble pas avoir d'effet globalement, il est par contre une grande influence lorsque qu'on se limite à l'un ou l'autre sexe (listing III.10.3). Les effets opposés à la fois s'amoindrissent mutuellement et par ailleurs expliquent le fort effet d'interaction

```
> summary(aov(op2[sexe==1] ~
  factor(age[sexe==1])))
              Df Sum Sq Mean Sq F value Pr(>F)
age[sexe==1] 2 30.519 15.259  23.211  2.5e-06 ***
Residuals   24 15.7778 0.6574

> summary(aov(op2[sexe==2] ~ fac-
tor(age[sexe==2])))
              Df Sum Sq Mean Sq F value Pr(>F)
age[sexe==2] 2 14.889 7.444   8.463   0.001654 **
Residuals   24 21.111 0.8796
```

**listing III.10.3**. Effet âge selon le sexe

De même, si le sexe ne semble pas avoir d'effet globalement, il est par contre une grande influence lorsqu'on se limite aux classes d'âge 1 et 3 (listing III.10.4). Les effets opposés s'annulent mutuellement. Par ailleurs, ils expliquent le fort effet d'interaction.

Le cas étudié ici est le plus classique (modèle fixe à deux dimensions). L'analyse de la variance peut prendre en compte d'autres modèles en distinguant les facettes croisées (auquel cas, on utilise le terme de dimension) et les facettes nichées. Par ailleurs, les facettes peuvent être fixes (les modalités épuisent toutes les valeurs possibles) ou aléatoires (les modalités considérées appartiennent à l'ensemble des valeurs possibles).

```
> summary(aov(op2[age==1] ~
factor(sexe[age==1])))
            Df Sum Sq Mean Sq F value Pr(>F)
sexe[age==1] 1 16.056 16.056  20.643  0.00033 ***
Residuals   16 12.4444 0.7778

> summary(aov(op2[age==2] ~
factor(sexe[age==2])))
            Df Sum Sq Mean Sq F value Pr(>F)
sexe[age==2] 1  2.000 2.000   2.5714  0.1284
Residuals   16 12.444 0.7778

> summary(aov(op2[age==3] ~
factor(sexe[age==3])))
            Df Sum Sq Mean Sq F value Pr(>F)
sexe[age==3] 1 24.50  24.50   32.667  3.2e-05 ***
Residuals   16  12.00  0.75
```

**listing III.10.4**. Effet sexe selon l'âge

La théorie de la généralisabilité regroupe la variance en trois parties : la variance de différentiation (la différence que l'on veut étudier, par exemple l'effet sexe), la variance d'instrumentation (les effets dus à d'autres variables identifiées, par exemple l'âge), la variance de généralisation (« l'erreur », c'est-à-dire la variabilité entre individus, elle regroupe ici l'interaction et la variance dans les groupes).

**Complément 1 : Les hypothèses d'artefact de l'ANOVA**

Il y a plusieurs approches de l'analyse de variance. Un modèle souvent présenté est celui du « modèle linéaire ». Les hypothèses qu'il sous-tend dans le cadre du paradigme population / échantillon sont les suivantes :

- Indépendance des observations : en particulier les groupes ne doivent pas être appariés.

- Normalité : la distribution du terme « d'erreur » (erreur de généralisabilité) suit une loi normale.

- Homogénéité de la variance : la variance est la même dans tous les groupes.

**Complément 2 : La formule fondamentale de l'analyse de variance**

Les valeurs de $X$ dans $G_i$ ($i$=1,..$k$) sont notées $x_{ij}$ ($j$=1,…$n_i$).

$$SST(X) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - m(X))^2$$

$$SST(X) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - m_i(X) + m_i(X) - m(X))^2$$

$$SST(X) = \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}(x_{ij} - m_i(X))^2 + 2\sum_{j=1}^{n_i}(x_{ij} - m_i(X))(m_i(X) - m(X)) + \sum_{j=1}^{n_i}(m_i(X) - m(X))^2\right]$$

$$SST(X) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - m_i(X))^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - m_i(X))(m_i(X) - m(X)) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(m_i(X) - m(X))^2$$

$$SST(X) = \sum_{i=1}^{k}SS_i(X) + \sum_{i=1}^{k}n_i(m_i(X) - m(X))^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - m_i(X))(m_i(X) - m(X))$$

$$SST(X) = SSd(X) + SSe(X) +$$

$$2\sum_{i=1}^{k}(m_i(X) - m(X))\sum_{j=1}^{n_i}(x_{ij} - m_i(X))$$

$$SST(X) = SSd(X) + SSe(X) +$$

$$2\sum_{i=1}^{k}(m_i(X) - m(X))(n_i m_i(X) - n_i m_i(X)) =$$

$$SSd(X) + SSe(X)$$

### Complément 3 : Exemple de calcul avec des groupes d'effectifs inégaux

On ne prend que les 50 premières observations du questionnaire. On peut constater qu'une cellule contient 5 individus et toutes les autres 9 (listing III.10.5).

```
> quest <- quest[1:50,]
> attach(quest)
> table(sexe,age)
age
sexe 1 2 3
   1 9 9 9
   2 9 9 5
```

**listing III.10.5**. Effectifs selon les variables sexe et age pour les 50 premières observations de la base quest.

Les résultats délivrés par la procédure `aov` de **R** dépendent alors de l'ordre des facteurs (listing III.10.6).

```
> summary(aov(op2 ~ factor(sexe) * factor(age),
quest[1:50,]))
            Df Sum Sq Mean Sq F value Pr(>F)
sexe         1  0.149  0.149   0.1922  0.6632
age          2  6.614  3.307   4.2686  0.0202*
sexe:age     2 30.028 15.014  19.3793  9e-07 ***
Residual    44 34.089  0.775

> summary(aov(op2 ~ factor(age)*factor(sexe),
quest[1:50,]))
            Df Sum Sq Mean Sq F value Pr(>F)
age          2  6.721  3.361   4.3377  0.0191*
sexe         1  0.042  0.042   0.0539  0.8175
age:sexe     2 30.028  15.014 19.3793  9.e-07 ***
Residual    44 34.089   0.775
```

**listing III.10.6**. Analyses de variance avec modification de l'ordre des facteurs

Un calcul « à la main » mène aux résultats présentés ci-dessous. On peut tout d'abord vérifier que : $SST = SSd+SSe$ . A partir des sommes des carrés des lignes et des colonne, on peut déduire la part due à l'interaction : $SSi = SSe-SSl–SSc$ = 29.92. Lorsque le calcul se fait directement pour obtenir $SSi2$, on obtient une valeur légèrement plus élevée (30.06). Le fait que $SSi < SSi2$ est une propriété générale.

- Moyenne générale[1] : `M = 2.68`
- Somme des carrés totale[2] : `SST = 70.88`
- Somme des carrés dans les groupes[3] : `SSd = 34.09`
- Somme des carrés entre les groupes[4] : `SSe = 36.79`
- Somme des carrés entre lignes[5] : `SSl = 0.149`
- Somme des carrés entre colonnes[6] : `SSc = 6.72`
- Calcul direct de la somme des carrés d'interaction[7] : `SSi2 = 30.06`

[1] `mean(op2)`

[2] `sum((op2-M)^2)`

[3] $\sum\sum (op2[sexe == i \;\&\; age == j] - M_{ij})^2$

[4] $\sum\sum n_{ij}(M_{ij} - M)^2$

[5] $\sum n_{i\bullet}(M_{i\bullet} - M)^2$

[6] $\sum n_{\bullet j}(M_{\bullet j} - M)^2$

[7] $\sum\sum n_{ij}(M_{ij} - M_{i\bullet} - M_{\bullet j} + M)^2$

## Chapitre III.11. Distributions « exacte » et approchée du coefficient *S*

La comparaison s'effectue sur un exemple sans ex-aequo de paramètres suivants :

$n = 10$, $t_1 = 8$, $t_2 = 2$. Le nombre de tableaux possibles est 45, chacun de multiplicité 1. La distribution exacte du *S* est la suivante (à compléter par symétrie pour les valeurs négatives de S) :

| S: | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| Dist. effectifs | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 1 | 1 |

Pour la distribution approchée, la valeur de $\sigma_S$ est obtenue par la formule : $t_1 * t_2 * (n+1) / 3$ . On trouve $\sigma_S = 7.66$.

La distribution exacte (densité de probabilité) associée (« hauteur » des rectangles de l'histogramme) est donnée par les fréquences relatives divisées par la mesure de l'intervalle existant entre deux valeurs successives de *S*. Par exemple, à la valeur 0 de *S* d'effectif 5 est associé le rectangle dont la base est l'intervalle [–1;1] et l'aire 5/45 (45 est la somme des effectifs). Sa « hauteur » vaut donc 5/90 = 0.056.

La distribution approchée est donnée par la formule :

$$\exp(-x^2/2\sigma^2)/ \sigma \sqrt{(2\pi)} .$$

| S: | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| d. exacte: | .056 | .044 | .044 | .033 | .033 | .022 | .022 | .011 | .011 |
| d. appr. : | .052 | .051 | .046 | .038 | .030 | .022 | .016 | .010 | .007 |

On constate que les valeurs sont semblables mais que la distribution approchée ne respecte évidemment pas les paliers de la distribution exacte.

## Chapitre III.12. Etude d'une interaction à l'aide du coefficient δ

On part d'une situation simple donnée par le tableau III.12.1.

| niv. Op. | | NC | I | C | Total |
|---|---|---|---|---|---|
| Garçons | F | 1 | 1 | 1 | 3 |
| | D | 1 | 1 | 0 | 2 |
| **Total** | | **2** | **2** | **1** | **5** |
| Filles | F | 1 | 1 | 0 | 2 |
| | D | 0 | 1 | 1 | 2 |
| **Total** | | **1** | **2** | **1** | **4** |
| Total | F | 2 | 2 | 1 | 5 |
| Total | D | 1 | 2 | 1 | 4 |
| **Grand total** | | **3** | **4** | **2** | **9** |

**tableau III.12.1**. Les données de l'étude

On peut calculer: $S_{\text{garçon}} = -2$ ; $\delta_{\text{garçon}} = -1/3$ ; $S_{\text{fille}} = 3$ ; $\delta_{\text{fille}} = 3/4$.

La différence $\Delta_{\text{delta}} = \delta_{\text{fille}} - \delta_{\text{garçon}} = 13/12 = 1.08$ est-elle « significative » ?

Les contraintes d'invariance des sommes marginales sont :

1. Sommes des lignes invariantes : cela représente trois contraintes (le total est fixé) naturelles, les groupes étant constitués.

2. Sommes des colonnes invariantes (contrainte faible): on suppose l'invariance du nombre d'individus pour chaque niveau opératoire (deux contraintes).

2'. Somme des colonnes invariantes pour chaque sous-groupe (contrainte forte): il y a invariance du nombre d'individus de chaque sexe et dans chaque condition expérimentale pour chacun des trois niveaux opératoires (4 contraintes).

On peut considérer trois cas de calculs exacts.

*Cas 1 : Contraintes maximales*

Il y a quatre *patterns* qui satisfont aux 8 contraintes 1) et 2') (tableau III.12.2).

La probabilité d'observer un effet aussi important est, dans ce cas, assez élevée, légèrement supérieure à 0.5 : $p = 8/15$.

| | | | | |
|---|---|---|---|---|
| Garçons F | 1 2 0 | 1 1 1 | 2 0 1 | 2 1 0 |
| Garçons D | 1 0 1 | 1 1 0 | 0 2 0 | 0 1 1 |
| delta | 1/6 | -1/3 | 1/3 | 5/6 |
| multiplicité | 2 | 4 | 1 | 2 |
| Filles F | 1 0 1 | 1 1 0 | 0 2 0 | 0 1 1 |
| Filles D | 0 2 0 | 0 1 1 | 1 0 1 | 1 1 0 |
| delta | 0 | 3/4 | 0 | -3/4 |
| multiplicité | 1 | 2 | 1 | 2 |
| $\Delta_{delta}$ | -1/6 | 13/12 | -1/3 | -19/12 |
| multiplicité | 2 | 8 | 1 | 4 |

**tableau III.12.2**. Les cas possibles sous contraintes 1) et 2')

*Cas 2 : Contraintes intermédiaires*

Si l'on remplace la contrainte 2') par 2), on trouve 109 *patterns* possibles. En comptant les multiplicités le nombre s'élève à 660 dont 55 ont une valeur de $\Delta_{delta}$ égale ou supérieure à 13/12. La p-value vaut 0.083.

*Cas 3 : Contraintes minimales*

Finalement, si on ne tient compte que de la contrainte 1), il y a 2160 *patterns* et 18900 en tenant compte de la multiplicité. La valeur de $\Delta_{delta}$ est égale ou supérieure à 13/12 dans 987 cas d'où une p–value de 0.052. Ce cas est celui qui correspond le mieux à une attribution aléatoire des étiquettes NC, I et C tout en paraissant extrême[1].

[1] La bibliothèque IPErad fournit les outils ad hoc pour effectuer ces calculs. La liste des tables pour le cas 1 est lstIII.12. Pour les cas 2 et 3, elle est générée respectivement par cree.pattern(c(3,2,2,2),c(3,4,2)) et cree.pattern2(c(3,2,2,2),3). La fonction fIII.12 prend une de ces listes en paramètre et renvoie la liste des multiplicités et des différences de delta. La fonction analIII.12 prend cette nouvelle liste en paramètre de même qu'une valeur *a* et renvoie la somme des multiplicités dont la différence de delta correspondante est supérieure ou égale à *a*.

## Chapitre III.13. Quelques propriétés du coefficient *S*

Le calcul du coefficient *S* et la vérification de certaines de ses propriétés peuvent être facilités sur la base d'une définition intrinsèque.

On considère V = $R^n$ espace vectoriel des suites de *n* nombres réels. Les éléments de cet espace seront donnés par leurs composantes: $v = (v_i)_{i=1,n}$

Exemple: $n = 4$, $v = (3, 4, 5, 6)$ ; $u = (1, 2, 3, 4)$

On définit un produit entre deux éléments *v* et *w* de cet espace V par : $v * w = \sum v_i \left( \sum_{j>i} w_j - \sum_{j<i} w_j \right)$

Ce produit jouit des propriétés suivantes :

1. anti-commutativité : $v*w = -w*v$ ;

2. homogénéité : $(\alpha\, v)*w = \alpha\, (v*w) = v*(\alpha\, w)$ où $\alpha$ est un nombre réel ;

3. additivité : $(v+v')*w = v*w + v*w'$ ; $v*(w+w') = v*w + v*w'$.

Cela s'exprime en disant que le produit est une forme bilinéaire antisymétrique sur $R^n$. Ce produit peut être donné par un produit matriciel :

$$v * w = vFw^t \quad avec \quad F = \begin{pmatrix} 0 & 1 & 1 & ... & 1 \\ -1 & 0 & 1 & ... & 1 \\ -1 & -1 & 0 & ... & 1 \\ ... & ... & ... & ... & ... \\ -1 & -1 & -1 & ... & 0 \end{pmatrix}$$

Le membre de droite représente le produit matriciel du vecteur ligne *v*, de la matrice *F* et du vecteur colonne $w^t$. La liste des propriétés peut être complétée :

4. $v*v = 0$ découle de l'antisymétrie.

5. On définit pour $v = (v_i)_{i=1,n}$ la permutation $p(v) = (v_{n-i})_{i=1,n}$. Exemple : $p(1, 2, 3, 4) = (4, 3, 2, 1)$. Avec cette notation, on a : $p(v)*p(w) = -v*w$.

6. Soit $T$ un tableau de deux lignes $v$ et $w$. $v*w$ correspond au coefficient $S$ noté ici $S(T)$.

7. Soient $T$ un tableau, $v$ sa première ligne, $T'$ le reste du tableau, $u(T')$ le vecteur constitué des totaux marginaux de $T'$ (colonne). On a :

$$S(T) = v*u(T') + S(T')$$

Cette relation permet de calculer la valeur de $S$ de proche en proche.

8. Soient $T$ un tableau, $a$ le premier élément de sa première ligne, $v$ le reste de sa première ligne, $T'$ le reste, $w$ la première colonne de $T'$, $T'''$ le reste. $T''$ est $T$ sans sa première colonne. $n(v)$ ou $n(T')$ représente la somme des éléments d'un vecteur ou d'un tableau.

$$T = \begin{pmatrix} a & v \\ w & T''' \end{pmatrix} = \begin{pmatrix} a & v \\ & T' \end{pmatrix} = \begin{pmatrix} a & \\ w & T'' \end{pmatrix}$$

$$S(T) = a\, n(T''') - n(v)n(w) + S(T') + S(T'')$$

9. $S$ change de signe lorsque l'on effectue une symétrie d'axe vertical du tableau. Cela découle de 5) et 7).

10. $S$ change de signe lorsque l'on effectue une symétrie d'axe horizontal du tableau. Cela découle de 1) et 7). Cette propriété permet de calculer le $S$ sur de petites calculatrices.

11. $S$ est invariant par transposition, cela découle de 8) et d'une démonstration directe dans le cas des tableaux de 2 lignes.

**Utilisation pour des plans factoriels**

Considérons un « plan factoriel » avec deux caractères indépendants dont le premier possède deux modalités (par exemple le caractère « sexe » avec les deux modalités : fille et garçon). Les modalités du deuxième caractère (par exemple le niveau socio-économique, Niv, de modalités : sup, moy et inf) sont hiérarchisées. De même que le caractère dépendant est ordinal (par exemple le niveau opératoire de modalités NC, I et C).

Un exemple d'un tel plan est donné par le tableau III.13.1. Dans ce cas, on démontre facilement que : $S_{\text{Total}} = S_{\text{Sexe}} + S_{+}$ avec :

- $S_{Total}$ est la valeur du $S$ calculée sur l'ensemble du tableau (18 cases).

- $S_{Sexe}$ est la valeur du $S$ calculée en regroupant les sous-groupes définis par Niv.

- $S_{+} = (S_{Niv \mid fille} + S_{Niv \mid garçon})$, où $S_{Niv \mid fille}$ est la valeur du $S$ calculée pour le groupe « fille ».

A noter que $S_{+}$ est inférieure à $S_{Niv}$ .

Avec l'hypothèse simplificatrice que $S_{Sexe}$ et $S_{+}$ sont indépendants, il est possible de décomposer la variance totale : $v_{Total} = v_{Sexe} + v_{+}$.

On peut considérer la distribution de $S_{+}$. Sa moyenne est nulle. Sa variance vaut : $v_{Total} - v_{Sexe}$ .

| Sexe | Niv | NC | I | C | Total |
|---|---|---|---|---|---|
| fille | sup | | | | |
| | moy | | | | |
| | inf | | | | |
| garçon | sup | | | | |
| | moy | | | | |
| | inf | | | | |
| Total | | | | | |

**tableau III.13.1**. Plan factoriel

On peut utiliser le schéma précédent pour le tableau arrangé de manière à inverser l'échelle socio-économique pour les garçons (tableau III.13.2).

Dans cette situation, le terme correspondant à $S_{+}$ est $S_{-} = (S_{Niv \mid fille} - S_{Niv \mid garçon})$ peut être pris comme une mesure de l'interaction entre les deux caractères lorsque les effectifs sont équilibrés. Cette mesure vaut $S_{-} = S'_{Total} - S_{Sexe}$ de moyenne nulle et de variance : $v_{-} = v'_{Total} - v_{Sexe}$ .

| Sexe | Niv | NC | I | C | Total | |
|---|---|---|---|---|---|---|
| Fille | sup | | | | $t_1$ | $tt_1$ |
| | moy | | | | $t_2$ | |
| | inf | | | | $t_3$ | |
| Garçon | inf | | | | $t_4$ | $tt_2$ |
| | moy | | | | $t_5$ | |
| | sup | | | | $t_6$ | |
| Total | | $u_1$ | | $u_2$ | n | |
| | | $u_3$ | | | | |

**tableau III.13.2**. Plan factoriel modifié

La variance exprimée peut aussi se calculer à partir des sommes marginales :

$$v_{-} = \frac{2\left(\sum tt_j^3 - \sum t_i^3\right) + 3\left(\sum tt_j^2 - \sum t_i^2\right)}{18} +$$

$$\frac{\left(\sum t_i^3 - \sum tt_j^3 - 3\left(\sum t_i^2 - \sum tt_j^2\right)\right)\left(\sum u_i^3 - 3\sum u_i^2 + 2n\right)}{9n(n-1)(n-2)} +$$

$$\frac{\left(\sum t_i^2 - \sum tt_j^2\right)\left(\sum u_i^2 - n\right)}{2n(n-1)}$$

**Reprise d'un exemple**

Cette technique offre une autre façon d'estimer l'interaction étudiée dans le chapitre II.8 sur la base du tableau II.26bis (tableau III.13.3). Le test sur la différence des valeurs δ conduisait aux valeurs suivantes : $\Delta\delta = 0.285$ ; $v = 0.083$ ; $z = 0.989$ ; $p = 0.161$.

La technique présentée ici mène aux valeurs suivantes :

- $S_{-} = S'_{\text{total}} - S_{\text{Sexe}} = 59$ ;
- $v_{-} = 3896.35$ ; $\sigma_{-} = 62.42$ ; $z = S_{-}/\sigma_{-} = 0.94$ ;
- $p = 0.17$

| | | NC | I | C | Total |
|---|---|---|---|---|---|
| F | Garçons | 2 | 4 | 6 | 12 |
| | Filles | 3 | 6 | 6 | 15 |
| D | Garçons | 3 | 7 | 10 | 20 |
| | Filles | 6 | 4 | 3 | 13 |
| Total | | 14 | 21 | 25 | 60 |

**tableau III.13.3**. Répartition des élèves selon la situation expérimentale, le sexe puis le niveau opératoire (reprise du tableau II.26bis)

## Chapitre III.14. La suite logistique

### Présentation

La suite logistique donnée par la relation $x_{n+1} = rx_n(1 - x_n)$ avec $r$ non nul inférieur ou égal à 4modélise de façon extrêmement élémentaire l'évolution d'une population. En prenant $r$ positif et $x_n$ le pourcentage des proies dans une population composée de proies et de prédateurs, la relation dit que la proportion des proies augmente en fonction du nombre de proies et diminue en fonction du nombre de prédateurs. Elle permet également de modéliser d'autres phénomènes où deux actions contradictoires sont menées de front.

Utilisée avec un temps continu sous la forme $\frac{dy}{dt} = ay\left(1 - \frac{y}{K}\right)$, la solution[1] est la courbe logistique (voir chapitre III.3), ce qui explique la nomenclature utilisée. Le fait assez mystérieux est que cette relation considérée du point de vue continu se résout avec une « sage » fonction alors que du point de vue discret la solution présente des phases chaotiques.

Son espace de phase est à une seule dimension. L'ensemble des points obtenus à partir d'une valeur $x_0$ constitue une trajectoire dans cet espace. Nous allons étudier ces trajectoires pour diverses valeurs du paramètre $r$ et de la valeur initiale.

**Cas 1 : $r = 2$**

- $x_0 < 0$ : Si les valeurs initiales sont négatives, les valeurs suivantes sont toutes négatives et divergent.

- $x_0 > 1$ : Par contre avec une valeur initiale supérieure à 1, les valeurs suivantes divergent positivement.

- $x_0 = 0$ : C'est un point fixe. Il est dit instable puisque à partir de valeurs proches de 0, la trajectoire s'en éloigne.

- $x_0 = 1$ : Conduit à 0.

- $x_0 > 0$ et $x_0 < 1$ : Cet intervalle constitue un bassin d'attraction pour l'attracteur ponctuel 0.5 solution non nulle de l'équation $x = 2x(1 - x) = 2x - 2x^2$ qui se simplifie en

[1] http://fr.wikipedia.org/wiki/Modèle_de_Verhulst (consulté : février 2009)

$x(1-2x)=0$. En prenant $x_0$ = 0.2, la valeur obtenue à la quatrième itération vaut approximativement 0.4998581. La figure III.14.1 illustre ce phénomène de convergence. La valeur limite est à lire sur l'axe horizontal.

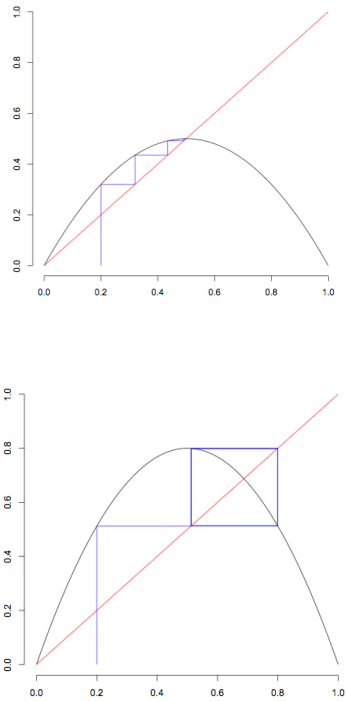



**fig III.14.1 et III.14.2**. Suite logistique pour r = 2 et r = 3.2

### Cas 2 : *r* = 3.2

Pour des valeurs initiales comprises entre 0 et 1, les valeurs de $x_n$ oscillent entre des valeurs proches de approximativement 0.5130 et 0.7995. La limite est un cycle de période 2. L'attracteur est constitué de 2 points (attracteur ponctuel périodique ou cyclique) (figure III.14.2). Ce fait peut s'expérimenter facilement. Pour une explication se basant sur des calculs simples, on peut se référer à Pilloud (1993). A noter que 0.6875, solution de $x = 3.2x(1-x)$, reste un point fixe, mais qui n'est pas stable.

**Cas 3 : $r = 3.5$**

Le même calcul conduit à une limite constituée d'un cycle constitué de 4 points (figure III.14.3). L'attracteur est à nouveau périodique. Les quatre valeurs sont approximativement : 0.5009, 0.8750, 0.3828 et 0.8269.

Pour des valeurs échelonnées jusqu'à $r_\infty \approx 3.5699456$, on trouve des attracteurs périodiques de périodes 8, 16, etc. Pour $r$ variant de $r_\infty$ à 4 la situation est plus complexe ; le système est « chaotique » dans le sens où il existe des attracteurs de toutes les périodes.

**Cas 4 : $r = 4$**

C'est un cas particulier d'un système chaotique. En particulier la valeur initiale $x_0 \approx 0.1169$ est 3-périodique. Ce qui implique par le théorème de Sarkovski qu'il existe des attracteurs pour n'importe quelle période entière.

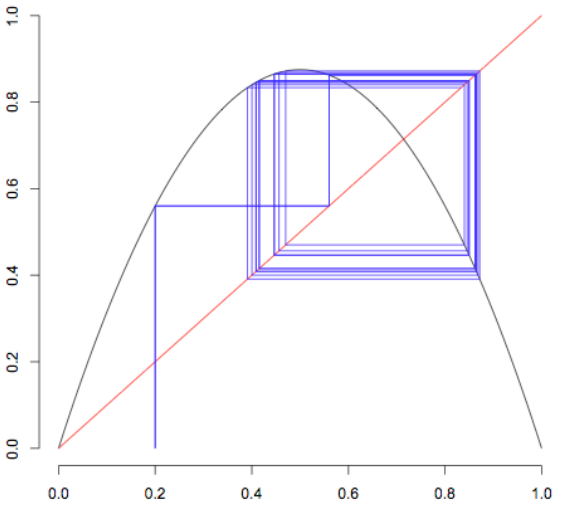



**fig III.14.3**. Suite logistique, $r = 3.5$

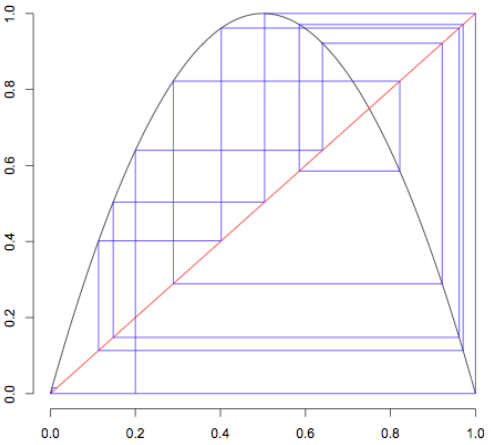



**fig III.14.4**. Suite logistique, $r = 4$

## Chapitre III.15. Construction de fractales entre ligne et surface

Cette annexe présente le processus de construction de fractales « pures » dans lequel l'aspect récurrent est bien visible

La première est connue sous le nom de triangle ou tapis de Sierpinski (figure III.15.1). Son processus de construction est le suivant. On considère tout d'abord un triangle. On considère ensuite le triangle inclus dont les sommets sont les milieux des trois côtés du triangle primitif. L'intérieur (sans la bordure du triangle lui-même) de ce triangle inclus est ôté. Le processus se répète dans les 4 triangles restants et ainsi de suite.

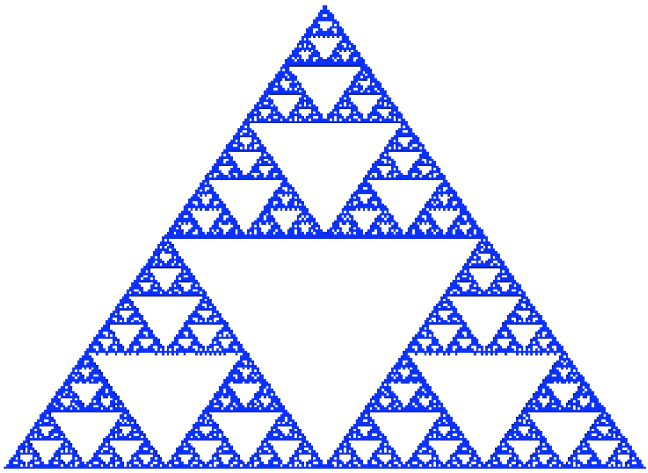
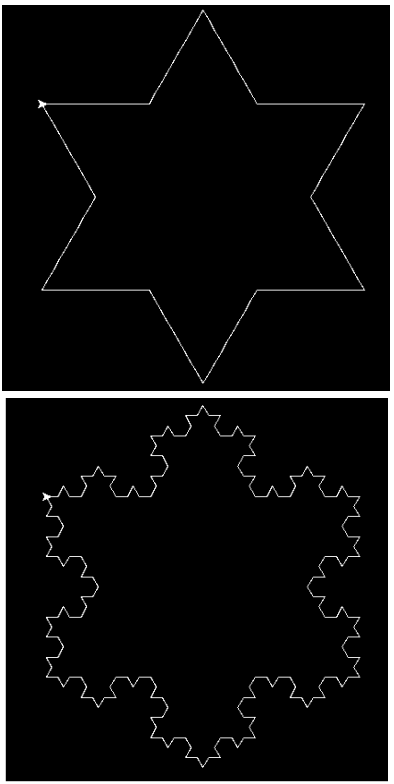


**fig III.15.1**. Triangle de Sierpinski de dimension fractale ln3/ln2 = 1.585

Un autre processus de construction est donné par l'exemple du flocon de von Koch (figure III.15.2) dont la construction du côté est donnée en code LOGO (listing III.15.1). On voit dans ce code la réutilisation de la procédure à l'intérieur de la procédure elle-même. Dans cette procédure, deux paramètres interviennent, la longueur (long) du côté et la profondeur (prof) de l'approximation. La dimension fractale du flocon vaut : ln4/ln3=1.2618.

```
to cote_flocon [long prof]
  if prof <= 0 [stop]
  elseif prof = 1 [fd long / 3 lt 60
                   fd long / 3 rt 120
                   fd long / 3 lt 60
                   fd long / 3  ]
           [cote_flocon long / 3  prof - 1 lt 60
            cote_flocon long / 3  prof - 1 rt 120
            cote_flocon long / 3  prof - 1 lt 60
            cote_flocon long / 3  prof – 1  ]
end
```

**listing III.15.1**. Procédure Logo pour la construction d'un côté du flocon de von Koch



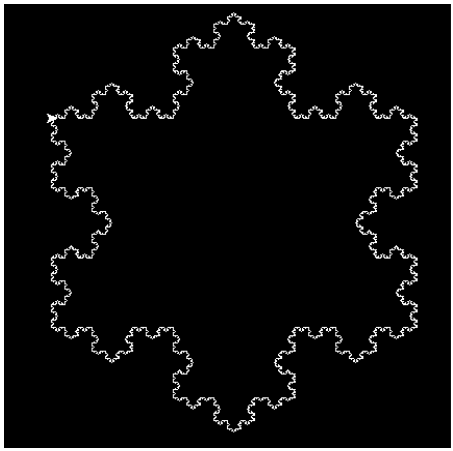


**fig III.15.2**. Le flocon de von Koch dessiné avec des profondeurs de 1, 3 et 5

## Chapitre III.16. L'entropie

Ce chapitre complète le chapitre II.9. Rappelons et précisons les trois types d'entropie généralement considérées.

**Entropie selon le lieu ou le mouvement**

Dans ce cas c'est la position des individus qui est prise en compte, leur qualité ne change pas comme dans l'exemple du café au lait issu de la physique déjà mentionné. Lors du mélange qui correspond à une augmentation du désordre (homogénéité), le café reste le café et le lait, le lait.

L'entropie selon la position est donnée par une formule « à la Boltzmann » où $n$ est le nombre d'entités réparties en $k$ groupes chacun d'effectif $n_i$ :

$$S = \frac{1}{n} \ln \frac{n!}{n_1!...n_k!} \quad \text{(II.9.I)}$$

Cette valeur atteint son maximum pour des valeurs de $n_i$ égales entre elles.

L'entropie par position peut aussi être calculée à l'aide de la formule de Shannon avec $p_i = n_i / n$ :

$$H = -\sum_{i=1}^{K} p_i \log_2 p_i \quad \text{(II.9.II)}$$

Ce coefficient peut s'interpréter comme un indice de dispersion. On notera $H_P$ la valeur obtenue par la même formule en remplaçant le logarithme de base deux par le logarithme naturel (le rapport entre ces deux valeurs vaut ln 2 = 0.693).

On peut vérifier que les valeurs obtenues par (II.9.I) et (II.9.II) sont à peu près proportionnelles. L'approche par (II.9.I) possède l'avantage de se baser sur un raisonnement lié à la position des entités. Elle table sur un état stable correspondant à l'état de probabilité maximale dans un comportement aléatoire. Par contre $H$ est plus facile à calculer.

$H$ ou $S$ (entropie de Shannon ou de Boltzmann) indique un manque d'information sur la position de $n$ éléments répartis dans $k$ groupes. Ces coefficients varient de $0$ à $\log_2 n$ pour $H$

et de 0 à $\frac{1}{n}\ln n!$ pour *S*. Pour un nombre *k* de groupes donné, *H* varie de *0* à $\log_2 k$ et *S* varie de *0* à $\frac{1}{n}\left[\ln n! - k \ln\left(\frac{n}{k}\right)!\right]$.

A titre d'exemple, les migrations peuvent être liées à ce modèle. Dans le cas de la localisation des habitants d'une ville, cet aspect de position peut entrer en conflit avec d'autres critères (distance au centre par exemple) qui seront examinés ultérieurement.

Le tableau III.16.1 présente les valeurs des coefficients *H*, $H_P$ et *S* pour quelques répartitions de 6 individus en 1, 2 ou 3 groupes. Le cas 1 présente l'entropie maximum (c'est dans ce cas que l'information concernant la localisation de chaque individu est la plus faible).

| Cas | Nombre d'individus | | | *H* | $H_P$ | *S* |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 1.59 | 1.10 | 0.75 |
| 2 | 1 | 2 | 3 | 1.46 | 1.01 | 0.68 |
| 3 | 1 | 4 | 1 | 1.25 | 0.87 | 0.57 |
| 4a | 0 | 2 | 4 | 0.92 | 0.64 | 0.45 |
| 4b | 0 | 0 | 6 | 0 | 0 | 0 |

**tableau III.16.1**. Différentes répartitions de 6 individus en 3 groupes avec les entropies de Shannon (*H* et $H_P$) et de Boltzmann (*S*) (reprise du tableau II.29)

**Entropie selon la qualité**

Les phénomènes correspondant à ce cas sont ceux où des groupes d'individus sont caractérisés par diverses qualités sans valeur hiérarchique et que l'on assiste à une homogénéisation par altération (les individus se modifient). Parmi les phénomènes qui correspondent à ce cas figure ceux de sociabilité ou d'imitation. On prend comme mesure du degré d'homogénéité une valeur proportionnel à la probabilité que deux individus tirés au hasard appartiennent au même groupe. Ce degré d'homogénéité vaut :

$$H_H = \sum_{i=1}^{k} \frac{n_i}{n} \frac{n_i - 1}{n - 1} = \frac{n}{n-1}\left( \sum_{i=1}^{k} p_i^2 - \frac{1}{n} \right)$$ , ou approximativement :

$$H_H \cong H'_H = \sum_{i=1}^{k} p_i^2$$

Avec $n$ nombre total d'individus, $k$ nombre de groupes et $p_i = n_i / n$ où $n_i$ est le nombre d'individus du groupe $i$.

Cette valeur est minimum (et vaut $1 / k$ avec $p_i = 1 / k$) lorsque tous les groupes comptent un nombre identique d'individus et d'autant plus que le nombre de groupes est grand. Il tend à diminuer lorsque les groupes sont de tailles différentes.

| Cas | Nombre d'individus | | | $H_H$ | $H'_H$ | $S$ |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 0.20 | 0.33 | 0.75 |
| 2 | 1 | 2 | 3 | 0.27 | 0.39 | 0.68 |
| 3 | 1 | 4 | 1 | 0.40 | 0.50 | 0.57 |
| 4a | 0 | 2 | 4 | 0.47 | 0.56 | 0.45 |
| 4b | 0 | 0 | 6 | 1 | 1 | 0 |

**tableau III.16.2**. Différentes répartitions de 6 individus en 3 groupes avec les entropies selon la qualité $H_H$ ($H'_H$) et de Boltzmann (S) (reprise de II.30)

En définitive, $H_H$ (entropie de variété)[1] indique une homogénéisation par assimilation d'un groupe par un autre ou le métissage. Pour $n$ individus, ce coefficient varie de *0* ($n$ groupes de 1 individu) à 1 (1 groupe de $n$ individus). Si on maintient une contrainte de $k$ groupes, ce coefficient varie de $\frac{n}{n-1}\left(\frac{1}{k} - \frac{1}{n}\right)$ (tous les groupes ont un nombre identique d'individus) à $\frac{1}{n-1}\left(\frac{(n-k)^2 + 2n - k}{n} - 1\right)$ (tous les groupes

[1] Le symbole $H_H$ est utilisé pour signaler l'horizontalité de la notion d'homogénéité alors que la hiérarchie sera considérée comme une dimension verticale.

comptent 1 seul individu sauf un groupe dans lequel se trouve le reste).

Le tableau III.16.2 compare les coefficients $S$ et $H_H$ qui varient de façon inverse comme cela était prévisible.

**Tendances à l'égalité et à l'inégalité**

Dans le cas où l'on attribue une certaine « utilité » (salaire, fortune, note scolaire, etc.) à chaque individu, une hiérarchie s'établit entre groupes qui sont constitués « selon la quantité ». A la tendance à l'homogénéité dans le cas de la qualité correspond naturellement une tendance à l'égalité. A cette tendance naturelle, il faut toutefois parfois opposé une tendance à l'inégalité qui dans le cas des systèmes humains serait « sociétale » (Forsé, 1989:146).

Le coefficient de Gini peut être utilisé pour mesurer l'état d'égalité. Si $X$ est la variable permettant d'effectuer la hiérarchie, on peut calculer la moyenne des différences[1] que l'on divise par $2\mu$ ($\mu$ moyenne de $X$) pour obtenir une coefficient compris entre 0 et 1:

$$G = \frac{1}{2n^2\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_j - x_i\right|$$

où $x_i$ est la valeur de $X$ associée à chacun des individus $s_i$ ($x_i=X(s_i)$) et $\mu$ la moyenne associée. Si l'on considère les $k$ groupes $g_k$ constitués chacun par les $n_k$ individus $s_i$ de même $x_i$ (noté $\xi_i$) on peut utiliser la formule :

$$G = \frac{1}{2n^2\mu}\sum_{i=1}^{k}\sum_{j=1}^{k}\left|\xi_j - \xi_i\right| n_j n_i$$

$G$ est minimum en cas d'égalité. Dans le cas du paradigme échantillon/population, l'estimateur non biaisé pour le coefficient de la population est donné par $\hat{G} = G\,n/(n-1)$. $\overline{G} = 1 - \hat{G}$ est maximum en cas d'égalité.

A noter que $G$ est une mesure de l'égalité relative, alors que $2\mu G$ donne une mesure absolue.

[1] http://mathworld.wolfram.com/GiniCoefficient.html (consulté : février 2008)

**Le paradoxe de Simmel**

Ce paradoxe provient du fait qu'il n'y a pas forcément coïncidence des minima et des maxima des indices d'homogénéité et d'égalité (Forsé, 1989:152). Une société, par exemple, peut-être très inégalitaire et en même temps très homogène lorsque la richesse est répartie parmi un petit groupe et que le grand reste est également pauvre. Au contraire une société peut-être très hétérogène en même temps qu'égalitaire.

Comme exemple, considérons 6 individus répartis en trois groupes. En vue d'utiliser le coefficient $G$, on convient d'une variable « utilité » (salaire, note, nombres d'objets, etc.) attribuée à chaque individu. Elle vaut pour les trois groupes respectivement 3, 2 et 1. Le tableau III.16.3 propose différentes répartitions selon la qualité des 6 individus.

Les principales constatations sont :

- La variation des coefficients $H_H$ (et $H'_H$) qui ne dépendent pas de l'utilité ne sont pas sujets aux mêmes variations que $G$ comme de bien entendu.

- Du point de vue de l'utilité, le cas 3 ne considère que deux groupes de respectivement 2 et 4 individus. Il en va de même pour le cas 4a. Il n'y a qu'un seul groupe dans le cas 4b.

- Le cas 3 est celui qui maximise $H_H$ et $\overline{G}$, il présente en quelque sorte un compromis entre homogénéité et égalité.

| Cas | Nombre d'individus | | | Ut. tot | $H_H$ ($H'_H$) | $G$ ($\hat{G}$) | $\overline{G}$ |
|---|---|---|---|---|---|---|---|
| | Gr. 1 (3) | Gr. 2 (2) | Gr. 3 (1) | | | | |
| 1 | 2 | 2 | 2 | 12 | 0.20 (0.33) | 0.22 (0.27) | 0.73 |
| 2 | 1 | 2 | 3 | 10 | 0.27 (0.39) | 0.23 (0.28) | 0.72 |
| 3 | 1 | 4 | 1 | 12 | 0.40 (0.50) | 0.17 (0.20) | 0.80 |
| 4a | 0 | 2 | 4 | 8 | 0.47 (0.56) | 0.17 (0.20) | 0.80 |
| 4b | 0 | 0 | 6 | 6 | 1 | 0 | 1 |

**tableau III.16.3**. Différentes répartitions de 6 individus en 3 groupes avec l'utilité fixée

Pour analyser plus spécifiquement l'aspect « quantité », le tableau III.16.4 considère trois groupes constitués respectivement de 1, 2 et 3 individus avec différentes utilités (les utilités sont celles attribuées à chaque individu).

Dans ce tableau, on constate :

- Les coefficients $H$ et $H_H$ constants puisqu'ils ne dépendent que du nombre d'individus de chaque groupe.

- $G$ varie ce qui montre que ce coefficient est bien indépendant des deux autres.

- Pour $G$, le cas 5 est de fait qu'un seul groupe du point de vue de l'utilité. Les cas 7 et 8 sont constitués de deux groupes. Toutefois les valeurs de $G$ ne dépendent pas du découpage en groupes.

| Cas | Utilité | | | Ut. Tot. | $H$ | $H_H$ | $G$ ($\hat{G}$) |
|---|---|---|---|---|---|---|---|
| | Gr. 1 (1) | Gr. 2 (2) | Gr. 3 (3) | | | | |
| 5 | 2 | 2 | 2 | 12 | 1.46 | 0.27 | 0 |
| 6 | 3 | 2 | 1 | 10 | 1.46 | 0.27 | 0.23 (0.28) |
| 7 | 1 | 4 | 1 | 12 | 1.46 | 0.27 | 0.33 (0.40) |
| 8 | 6 | 0 | 0 | 6 | 1.46 | 0.27 | 0.83 (1) |

**tableau III.16.4**. Diverses utilités attribuées à 6 individus répartis en trois groupes

**Entropie sur l'utilité**

Le raisonnement concernant les états stables (répartition la plus probable) appliqué aux utilités suggère de considérer le coefficient :

$$H_V = \frac{1}{x} \ln \frac{x!}{(n_1 x_1)!...(n_k x_k)!}$$

où $x = \sum_{i=1}^{k} n_i x_i$ ($x_i$ est l'utilité attribuée à chaque individu du groupe d'indice $i$), avec toujours : $n = \sum_{i=1}^{k} n_i$.

Le tableau III.16.5 reprend les cas du tableau III.16.4 avec cette fois les valeurs de $G$, $\overline{G}$ et $H_V$. Il suggère que ces deux derniers coefficients donnent des informations comparables.

| Cas | Utilité | | | Ut. tot. | $G$ | $\overline{G}$ | $H_V$ |
|---|---|---|---|---|---|---|---|
| | Gr. 1 (1) | Gr. 2 (2) | Gr. 3 (3) | | | | |
| 5 | 2 | 2 | 2 | 12 | 0 | 1 | 0.79 |
| 6 | 3 | 2 | 1 | 10 | 0.23 | 0.72 | 0.83 |
| 7 | 1 | 4 | 1 | 12 | 0.33 | 0.60 | 0.63 |
| 8 | 6 | 0 | 0 | 6 | 0.83 | 0 | 0 |

**tableau III.16.5**. Comparaison de $G, \hat{G}$ et $H_V$ sur des groupes comprenant respectivement 1, 2, et 3 individus

La valeur maximale de $H_V$ avec les contraintes $n$ et $x$ fixés implique que toutes les valeurs des produits $n_i x_i$ sont égales[1]. En d'autres termes, les $n_i$ étant fixés, on a : $x_i = 1 / n_i$. Ce fait s'observe dans le tableau III.16.6 qui considère systématiquement tous les cas possibles de répartitions de 6 utilités entre des groupes comprenant 1, 2 et 3 individus. Le tableau donne l'utilité attribuée à chaque individu.

| Cas | Utilité | | | Ut. tot. | $G$ | $\overline{G}$ | $H_V$ |
|---|---|---|---|---|---|---|---|
| | Gr. 1 (1) | Gr. 2 (2) | Gr. 3 (3) | | | | |
| 8 | 6 | 0 | 0 | 6 | 0.83 | 0 | 0 |
| 9 | 0 | 0 | 2 | 6 | 0.50 | 0.40 | 0 |
| 10 | 0 | 3 | 0 | 6 | 0.67 | 0.20 | 0 |
| 10 | 2 | 2 | 0 | 6 | 0.50 | 0.40 | 0.45 |
| 11 | 3 | 0 | 1 | 6 | 0.50 | 0.40 | 0.50 |
| 12 | 1 | 1 | 1 | 6 | 0 | 1 | 0.68 |
| 13 | 2 | 1 | 2/3 | 6 | 0.22 | 0.73 | 0.80 |

**tableau III.16.6**. 6 utilités réparties en trois groupes comprenant respectivement 1, 2 et 3 individus

L'examen du tableau III.16.6 permet quelques remarques additionnelles.

- Les coefficients $\overline{G}$ et $H_V$ présentent une tendance commune bien qu'ils se basent sur des aspects différents. $\overline{G}$ tient

[1] La simplification de la valeur à maximiser conduit Forsé à une mauvaise approximation de la solution notamment pour le cas discret.

compte des différences, ce que $H_V$ ne prend pas en compte, ne s'intéressant qu'à la disposition. Pour un nombre égal d'utilités distribuées à chaque individu, $\overline{G}$ est maximum alors que ce n'est pas forcément le cas pour $H_V$. Par contre $H_V$ est nul lorsque toute l'utilité attribuée à un seul groupe alors que $\overline{G}$ atteint un minimum local qui peut varier.

- $H_V$ (entropie sur la répartition d'utilités) table sur un état stable correspondant à l'état de répartition des utilités le plus probable, dans un comportement aléatoire, dans un certain nombre de groupes, sachant que la répartition se fait sur la base d'un nombre donné d'individus par groupe, chacun recevant le même nombre d'utilités. Cette valeur est nulle lorsque toutes les utilités sont attribuées à un seul groupe. Elle est maximale lorsque les utilités sont réparties de façon égale dans chaque groupe. Dans le cas de groupes inégaux, l'entropie est maximale lors d'une distribution inégale, mais pas trop inégale. Cela signifie que lorsque les groupes sont de grandeurs différentes on attribue beaucoup à peu d'individus et réciproquement.

L'usage de ces différents coefficients pour qualifier l'état d'un système, mesurer son état de désordre et d'en repérer l'état de fragilité, demande évidemment de repérer le bon processus. Le calcul de plusieurs coefficients dans un même cas conduit peut conduire à des paradoxes, celui de Simmel est longuement étudié par Forsé (1989).

**Changement selon le lieu sous contraintes de la quantité**

Toutefois, la dynamique générale peut-être liée à la concurrence de différents mouvements caractérisés chacun par un indice entropique. Il est notamment possible de chercher le maximum pour les valeurs de $S$ et $H_H$ avec une condition liée à des répartitions d'utilités.

Dans le cas de $S$, sans contrainte, le système tend à l'équipartition (le nombre de groupes étant donné). Avec une contrainte liée aux utilités des aspects horizontaux et hiérarchiques se mêlent.

Il est relativement facile de quantifier quelques cas lorsque l'utilité totale est fixée constante (cas des ressources limitées). Dans ce cas, on peut considérer l'entropie selon la

position (changement de position dans un système hiérarchique). Les contraintes transforment le chaos entropique selon $S$ en une hiérarchie stable en rejoignant par là, selon Forsé, la logique des systèmes dissipatifs (Forsé, 1989:197). La position des habitants dans une ville, en considérant la distance à un centre d'intérêt relève du même processus (Forsé, 1989:225). Il est également possible d'étudier le modèle lorsque c'est $H_H$ qui est soumis à des contraintes.

*Maximisation d'entropie de lieu sous contrainte hiérarchique*

On suppose la distribution des $x_i$ donnée de même que $k$, le nombre de groupes. On utilisera $H_p$ comme mesure de l'entropie[1]. La valeur à maximiser est :

$$H_P = -\sum_{i=1,..k} p_i \ln p_i = -\frac{1}{n}\sum_{i=1,..k} n_i\left(\ln n_i - \ln n\right) = \ln n - \frac{1}{n}\sum_{i=1,..k} n_i \ln n_i$$

avec les contraintes : (1) $\sum n_i = n$ et (2) $\sum n_i x_i = x$.

La méthode des multiplicateurs de Lagrange demande de considérer :

$$L = \ln n - \frac{1}{n}\sum_{i=1,..k} n_i \ln n_i + \mu\left(n - \sum n_i\right) + \lambda\left(x - \sum n_i x_i\right)$$

La dérivée de cette expression par rapport à $n_i$ vaut :

$$\frac{dL}{dn_i} = -\frac{1}{n}\left(\ln n_i + 1\right) - \mu - \lambda x_i$$

Les extrêma sont donc donnés par la relation : $\ln n_i + n\mu + \lambda n x_i + 1 = 0$, ce qui conduit à :

$$n_i = \frac{1}{e^{n\mu+\lambda n x_i+1}} = \frac{1}{e^{n\mu+1}}\frac{1}{e^{\lambda n x_i}}$$

[1] Maximiser $S$ demande des approximations que l'on évite avec $H$. Pour simplifier on utile le logarithme naturel plutôt que le logarithme en base deux pour simplifier les expressions. Cela ne change pas les informations concernant le maximum.

par (1) on a : $n = \frac{1}{e^{n\mu+1}} \sum e^{-\lambda n x_i}$ et donc : $n_i = \frac{n e^{-\lambda n x_i}}{\sum e^{-\lambda n x_i}}$ ;

$f_i = \frac{n_i}{n} = \frac{e^{-\lambda n x_i}}{\sum e^{-\lambda n x_i}}$ (3)

par (2) on a : $\bar{x} = \frac{x}{n} = \sum f_i x_i = \frac{\sum x_i e^{-\lambda n x_i}}{\sum e^{-\lambda n x_i}}$ (4) .

Cette formule peut aussi s'écrire sous la forme symétrique :

$$\bar{x} \sum e^{-\lambda n x_i} = \sum x_i e^{-\lambda n x_i}$$

Résoudre (4) conduit à trouver λ puis $n_i$ à l'aide de (3). Il n'est pas aisé de procéder à une résolution générale. Cette résolution se fera dans deux cas. Puis une approximation générale sera faite en supposant le nombre de groupe allant vers l'infini avec $x_i$ prenant toutes les valeurs entières.

1) Résolution pour $n = 6$ ; $k = 3$ ; $x_i = i$ ($i = 1, 2, 3$) ; $x = 12$ ; $\bar{x} = 2$

Par (4) :

$$2 = \frac{1 \cdot e^{-6\lambda} + 2 \cdot e^{-12\lambda} + 3 \cdot e^{-18\lambda}}{e^{-6\lambda} + e^{-12\lambda} + e^{-18\lambda}} = \frac{1 + 2 \cdot e^{-6\lambda} + 3 \cdot e^{-12\lambda}}{1 + e^{-6\lambda} + e^{-12\lambda}}$$

$$2(1 + e^{-6\lambda} + e^{-12\lambda}) = 1 + 2 \cdot e^{-6\lambda} + 3 \cdot e^{-12\lambda}$$

$$1 + 2 \cdot e^{-6\lambda} + 2 \cdot e^{-12\lambda} = 2 \cdot e^{-6\lambda} + 3 \cdot e^{-12\lambda}$$

$$1 = e^{-12\lambda}$$

On en tire $\lambda = 0$ et par (3) ; $n_i = 2$

2) Résolution pour $n = 6$ ; $k = 3$ ; $x_1 = 1$ ; $x_2 = 2$ ; $x_3 = 4$ ; $x = 12$ ; $\bar{x} = 2$

Par (4) :

$$2 = \frac{1 \cdot e^{-6\lambda} + 2 \cdot e^{-12\lambda} + 4 \cdot e^{-24\lambda}}{e^{-6\lambda} + e^{-12\lambda} + e^{-24\lambda}} = \frac{1 + 2 \cdot e^{-6\lambda} + 4 \cdot e^{-18\lambda}}{1 + e^{-6\lambda} + e^{-18\lambda}}$$

$$2(1 + e^{-6\lambda} + e^{-18\lambda}) = 1 + 2 \cdot e^{-6\lambda} + 4 \cdot e^{-18\lambda}$$

$$1 = 2 \cdot e^{-12\lambda}$$

On tire $\lambda = -\frac{1}{12}\ln 0.5 = 0.0578$ et par (3) :

$n_3 = 1.14$ ; $n_2 = 1.62$ ; $n_1 = 3.24$ et en arrondissant : $n_3 = 1$ ; $n_2 = 2$ ; $n_1 = 3$.

3) Résolution pour $k$ et donc $n$ « grands » ; $x_1 = 1$, $x_2 = 2$, … $x_k = k$

On pose $a = e^{-\lambda n}$ et l'on remplace $x_j$ par $j$.

Le membre de droite de (3) devient : $a^j \Big/ \sum_{i=1}^{k} a^i$

Le membre de droite de (4) s'écrit : $\sum_{j=1}^{k} j a^j \Big/ \sum_{j=1}^{k} a^j$. Lorsque $a$ est inférieur à 1 (ce qui est le cas pour $n$ suffisamment grand) les séries convergent.

On a : $\sum_{j=1}^{\infty} a^j = \frac{a}{1-a}$ (série géométrique) puis en dérivant terme à terme :

$$\sum_{j=1}^{\infty} j a^{j-1} = \frac{1}{(1-a)^2} \text{ et donc } \sum_{j=1}^{\infty} j a^j = \frac{a}{(1-a)^2}$$

En définitive (4) devient :

$\bar{x} \cong \frac{1}{1-a} = \frac{1}{1-e^{-\lambda n}}$ , ce qui permet de déduire $e^{-\lambda n} = 1 - \frac{1}{\bar{x}} = \frac{\bar{x}-1}{\bar{x}}$

En définitive (3) devient :

$$f_j = a^{j-1}(1-a) = \left(1 - \frac{1}{\bar{x}}\right)^{j-1} \frac{1}{\bar{x}} = \frac{1}{\bar{x}-1}\left(1 - \frac{1}{\bar{x}}\right)^j$$

*Extrêmum de l'entropie de qualité sous contrainte hiérarchique*

La technique des multiplicateurs de Lagrange appliquée précédemment mène à trouver un minimum d'entropie de « qualité ». Le minimum indique qu'il n'est pas possible de

diversifier davantage les groupes. L'usage pratique de cette opération reste à préciser (recherche de solutions efficaces dans des processus d'innovation ?).

Par ailleurs, il s'agitait encore de caractériser les valeurs maximales qui doivent correspondre à des valeurs « frontières » (et non à des points d'extrêmum intérieurs au domaine).

On suppose la distribution des $x_i$ donnée, de même que $k$, le nombre de groupes. La valeur à « extrêmiser » est $H_H$ ou, ce qui revient au même, la valeur approchée qui s'obtient par transformation affine : $n^2 H_H \approx \sum_{i=1,..k} {n_i}^2$, avec des contraintes identiques : (1) $\sum n_i = n$ et (2) $\sum n_i x_i = x$.

Dans ce qui suit on note $\alpha = \sum x_i$ et $\beta = \sum {x_i}^2$ .

La méthode des multiplicateurs de Lagrange conduit à considérer :

$$L = \sum_{i=1,..k} {n_i}^2 + \mu\left(n - \sum n_i\right) + \lambda\left(x - \sum n_i x_i\right)$$

Et l'on calcule comme précédemment :

- Les dérivées partielles : $\frac{dL}{dn_i} = 2n_i - \mu - \lambda x_i = 0$ ;

- puis : $n_i = \frac{\mu + \lambda x_i}{2}$ ; $n = \sum n_i = \frac{k\mu + \lambda\alpha}{2}$ ; $\mu = \frac{2n - \lambda\alpha}{k}$ ;

- $n_i = \frac{2n + \lambda(kx_i - \alpha)}{2k}$ ;

- $x = \sum n_i x_i = \frac{1}{2k} \sum (2n + \lambda(kx_i - \alpha)) x_i = \frac{1}{2k} (2n\alpha + \lambda(k\beta - \alpha^2))$ ;

- $\lambda = \frac{2kx - 2n\alpha}{k\beta - \alpha^2}$ ;

- et finalement : $n_i = \frac{n}{k} + \frac{(kx - n\alpha)(kx_i - \alpha)}{k(k\beta - \alpha^2)}$.

**Applications**

1. $x = 12$ ; $n = 6$ ; $k = 3$ ; $x_i = i$

On peut calculer : $\alpha = 6$ ; $\beta = 14$ et l'on trouve finalement : $n_i = 2 + 0\ (3x_i - 6)$ ; $n_i = 2$ pour $i$=1,2,3

2. $x = 12$ ; $n = 6$ ; $k = 3$ ; $x_i = 1, 2, 4$

On peut calculer : $\alpha = 7$ ; $\beta = 21$ et l'on trouve finalement : $n_i = 2 - (1/7)\,(3x_i - 7)$ et donc :

$n_1 = 2 + (4/7) \approx 3$ ; $n_2 = 2 + (1/7) \approx 2$ ; $n_3 = 2 - (5/7) \approx 1$

3. $k = 6$ ; $n = 12$ ; $x_i = i$ $(i=1,.., 6)$

On peut calculer : $\alpha = n(n+1)/2 = 21$; $\beta = n(n+1)(2n+1)/6 = 91$

3a. $x=30 \Rightarrow n_1 = 4$ ; $n_2 = 3$ ; $n_3 = 2$ ; $n_4 = 2$ ; $n_5 = 1$ ; $n_6 = 0$

3b. $x=42 \Rightarrow n_1 = 2$ ; $n_2 = 2$ ; $n_3 = 2$ ; $n_4 = 2$ ; $n_5 = 2$ ; $n_6 = 2$

3c. $x=50 \Rightarrow n_1 = 1$ ; $n_2 = 1$ ; $n_3 = 2$ ; $n_4 = 2$ ; $n_5 = 3$ ; $n_6 = 3$

Dans ce cas, on observe que selon la valeur de la ressource totale, le profil des groupes n'est pas le même (les valeurs ont été arrondies).

Pour vérifier qu'il s'agit bien d'un minimum, le calcul de l'entropie donne pour le cas 3a : $H_H = 0.17$.

Avec $x = 30$, $(n_i) = (5, 2, 2, 1, 1, 1)$ ($k = 6$ ; $n = 12$), on trouve $H_H = 0.18$.

**Activités pour réfléchir**

*Egalité absolue et relative*

$G$ est une mesure de l'égalité relative, alors que $2\mu G$ est une mesure « absolue ». Examinez différents cas en situant chaque fois ces deux coefficients (diminution absolue de l'utilité de $x$ point pour chaque individu, diminution relative de moitié pour chaque individu, augmentation relative ou absolue, proportionnelle à l'utilité actuelle, etc.).

*Les démons*

Dans le modèle entropique le dispositif qui sous la forme d'apport d'information ou d'énergie permet de diminuer l'entropie (augmentation de l'ordre) est souvent appelé un « démon » en référence au démon de Maxwell qui a servi à stabiliser la notion d'entropie en thermodynamique. Le bibliothécaire qui remet de l'ordre dans sa bibliothèque est, dans le cadre de cette théorie, un démon !

Wikipédia[1] nous apprend qu'à la suite des travaux de Morton Grodzins, T. Schelling montre à quelles conditions un quartier où des blancs et des noirs sont mélangés peut devenir pratiquement noir même si ce n'est pas ce que souhaitaient ses habitants. Ce cas correspond à une diminution de l'entropie selon le lieu ou le mouvement grâce à l'apport d'information et/ou d'énergie.

Etudiez ce cas connu sous l'appellation de la ségrégation de Schelling ou du démon de Schelling.

Il est également possible d'en faire une simulation[2]. La règle est simple, on attribue à chaque individu un taux de tolérance qui représente une proportion d'habitants « différents » dans son quartier. Si la proportion de gens différents de lui dans son quartier est supérieure à ce taux, l'individu quitte le quartier et s'établit ailleurs, au hasard. Le résultat final dépend de la proportion de départ et du taux de tolérance. Il existe un point critique en deçà duquel les deux couleurs vont rester mélangées et au-delà duquel on assiste à une ségrégation.

[1] http://fr.wikipedia.org/wiki/Thomas_Schelling (consulté : décembre 2009).

[2] Pour une simulation voir la situation « pour réfléchir » liée à NetLogo (voir les activités pour réfléchir du chapitre I.8) ou consulter sur le Web, la page :
http://community.ofset.org/index.php/La_ségrégation_selon_Thomas_Schelling (consulté : décembre 2009).

## Chapitre III.17. Introduction au vocabulaire de la théorie des graphes

### Un brin d'histoire

Un articles, parmi les largement connus, d'Euler (1707-1783) a trait au problème des ponts de Königsberg (à l'époque ville de Prusse, actuellement Kaliningrad, enclave russe en Allemagne). Il a été présenté à l'Académie de St. Pétersbourg le 26 août 1735.

*Le problème*

La figure III.17.1 présente un plan de la ville de Königsberg traversée la rivière Pregolia qui entoure deux grandes îles reliées entre elles et aux deux rives par sept ponts.

Le problème des ponts de Königsberg est de relier les parties A, B, C, D de la ville en revenant à son point de départ en passant une fois et une seule sur chacun des sept ponts.

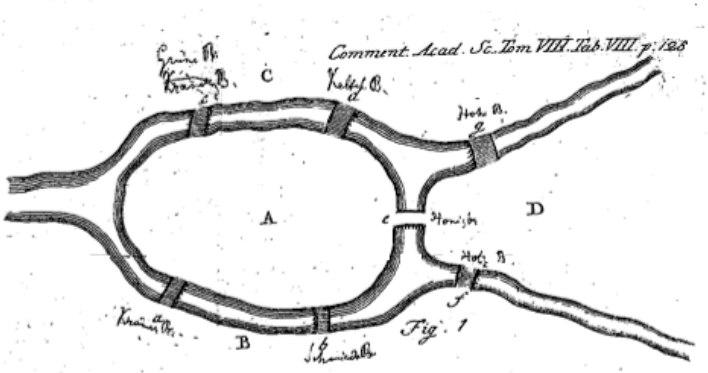



**fig III.17.1**. Illustration du problème des ponts de Königsberg tirée de l'article orignal (Euler, 1741)

Il a été résolu par Euler par la négative qui a donné en même temps des conditions suffisantes et nécessaires pour que cela soit possible en généralisant au nombre de sommets et d'arêtes.

L'article d'Euler Il est souvent cité comme l'article fondateur de la topologie et de la théorie des graphes. Cette dernière a pour but d'étudier des configurations « abstraites »

de sommets reliés entre eux, de dégager des propriétés et des lois. Dans ce cadre, le problème des ponts de Königsberg se réduit au schéma de la figure III.17.2.

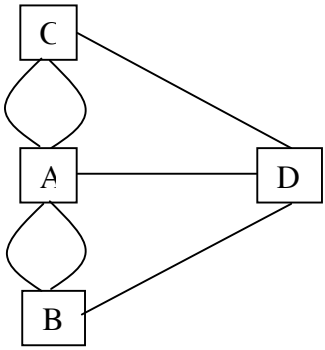



**fig III.17.2**. Le graphe « abstrait » du problème des ponts de Königsberg

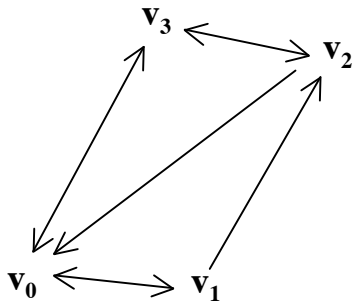
**Utilité de la théorie des graphes**

Il n'est pas difficile d'imaginer des situations qui se laissent modéliser par un graphe : situations concrètes de parcours ou de flux et situations plus abstraites (liens entre idées). Il est fait un large usage des techniques de la théorie dans l'analyse de la structure et de la dynamique des réseaux sociaux dont le sociogramme est un exemple historique. Cette théorie se prolonge dans plusieurs directions, topologie générale en mathématique, réseaux sémantiques en sciences cognitives, etc.

**Définition formelle**

Un graphe est donné par un ensemble de sommets V et d'une relation de V sur V (un ensemble de couples de sommets). Les sommets peuvent être représentés par des points du plan $v_1$, $v_2$, … et chaque couple $(v_1,v_2)$ par une flèche d'origine $v_1$ et de pointe $v_2$ (figure III.17.3).

Un graphe est dit orienté si l'on tient compte du sens des flèches, non orienté sinon. Pour traiter un graphe par des moyens mathématico-informatiques, on représente un graphe par sa matrice d'adjacence (figure III.17.3). Une valeur 1 au croisement de la ligne i et de la colonne j correspond à l'existence de l'arête $(v_i,v_j)$. Une matrice d'adjacence symétrique correspond à un graphe non orienté.




| | [0] | [1] | [2] | [3] |
|---|---|---|---|---|
| [0] | 0 | 1 | 0 | 1 |
| [1] | 1 | 0 | 1 | 0 |
| [2] | 1 | 0 | 0 | 1 |
| [3] | 1 | 0 | 1 | 0 |

**III.17.3**. Un graphe et sa matrice d'adjacence

Les concepts les plus souvent rencontrés sont les suivants :

- Graphe complet : c'est un graphe (non orienté) pour lequel une arête existe entre chaque sommet.

- Chemin : c'est une suite d'arêtes dans un graphe orienté de telle manière que l'origine d'une arête corresponde à l'extrémité terminale de la précédente. Un circuit est un chemin fermé (l'extrémité terminale de la dernière arête de la suite est l'origine de la première). Ces notions sont également valables pour un graphe non orienté.

- Connexité : un graphe non orienté est connexe est un graphe d'un seul tenant. Il y a toujours un chemin entre deux sommets quelconques. Lorsqu'un graphe n'est pas connexe, il peut se décomposer en composantes connexes. Pour un graphe orienté, la notion associée est plus délicate. On définit le « core » d'un sommet comme l'ensemble des sommets atteignables par un chemin depuis le sommet choisi et tels qu'il existe également un chemin de retour.

- Indices de structure : les indices classiques sont ceux de centralité sortante et rentrante (indiquent respectivement le nombre d'arêtes issues du sommet et le nombre d'arêtes qui aboutissent au sommet), compacité, linéarité (pour des définitions plus précises, voir Pochon & Favre, 2007).

- Eléments de structure : on peut distinguer des parties dans un graphe comme les étoiles, les cliques (sous-graphes complets), etc.

**Quelques outils disponibles dans R**

La « librairie » igraph permet, entre autres choses, de définir un graphe à partir de sa matrice d'adjacence (graph.adjacency) et de le représenter (plot). En utilisant la librairie tcltk, la fonction tkplot fournit une représentation que l'on peut ajuster « à la souris » (listing III.17.1 et figure III.17.4, `mat20` est une matrice d'adjacence se trouvant dans exGraph.dat à charger à l'aide de la commande `load`).

```
> library(igraph)
> library(tcltk)
> gr20 <- graph.adjacency(mat20, mode="dir")
> tkplot(gr20)
```

**listing III.17.1**. création d'un graphe à partir d'une matrice d'ordre 20.

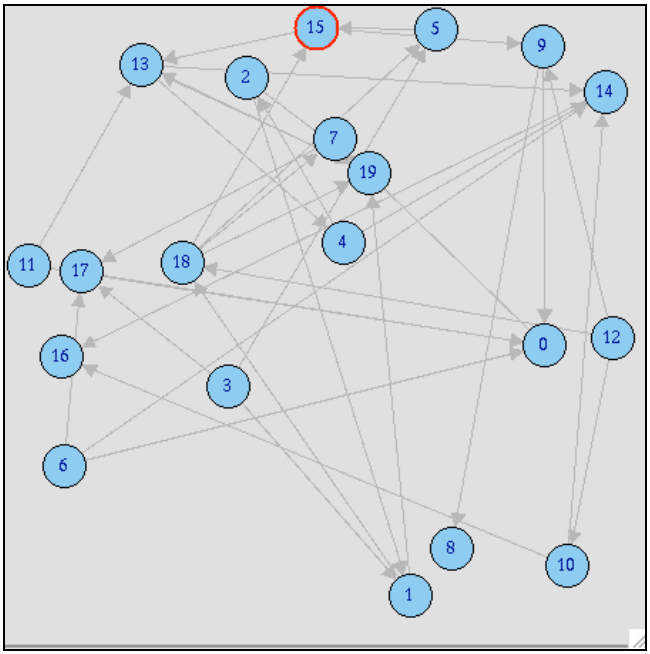



**fig III.17.4**. Le graphe correspondant au listing III.17.1

La bibliothèque IPErad propose des fonctions qui permettent de calculer les indices de centralités sortantes et rentrantes d'un graphe (listing III.17.2), de chercher la cardinalité des « cores » de chaque sommet (listing III.17.3) et encore de décomposer un graphe en « cores » (listing III.17.4 et figure III.17.5).

La commande mcore décompose la matrice du graphe en regroupant les éléments par « cores ».

```
> source("R-hptxt-info")
> centralite(mat20)
0   1   2   3   4   5   6   7   8   9   10  11  12  13
1.1 2.3 2.2 3.2 2.2 1.1 3.2 1.1 0.0 2.1 2.1 2.2 3.4 3.2
4.3 2.1 1.1 0.0 1.2 2.1 0.0 2.3 1.1 2.1 1.0 0.0 0.0 3.3
```

**listing III.17.2**. Indices de centralités sortantes (deuxième ligne) et entrantes (troisième ligne) des 14 premiers sommets (première ligne).

```
> core(mat20)
> 3 8 8 0 8 8 0 3 0 0 0 0 0 8 0 8 0 3 8 8
```

**listing III.17.3**. Cardinalité des « cores » des 20 sommets

```
> tkplot(grap.adjacency(mcore(mat20),
mode= "dir"))
```

**listing III.17.4**. Affichage des « cores » du graphe défini par mat20

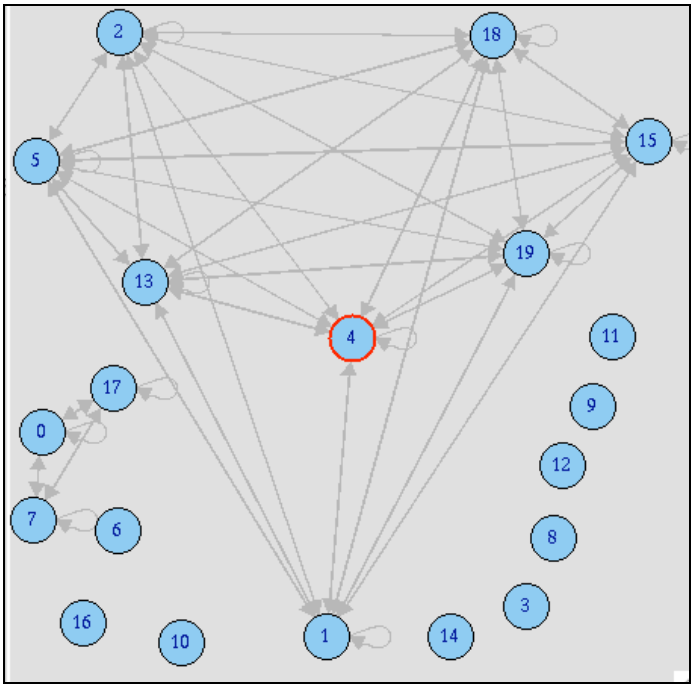



**fig III.17..5**. résultat de la commande du listing III.17.4. On observe les deux cores de respectivement 8 et 3 éléments et les 9 éléments isolés de ce point de vue

## Chapitre III.18. La « librairie » IPErad pour R

Cette annexe constitue une brève introduction à l'utilisation de **R**. Elle rassemble des éléments épars présentés dans le reste de l'ouvrage et détaille les outils disponibles dans la librairie IPErad. Pour les manipulations de base, on pourra aussi se référer à des exemples liés à des travaux pratiques en psychologie proposés par Pallier & Lalanne (2005).

### Installer R

Téléchargez et installez **R** à partir de l'adresse `http://www.r-project.org/` (choisissez sous CRAN mirror un site de téléchargement proche de chez vous, l'Université de Berne ou l'ETHZ pour la Suisse, le CICT à Toulouse ou le Département de biométrie et de biologie évolutionniste de l'université de Lyon pour la France, etc.).

Pour les manipulations, utilisez le jeu de données et la bibliothèque dont les dernières versions peuvent être téléchargées sur le site de l'Institut de psychologie et éducation de l'université de Neuchâtel[1].

Choisissez un répertoire de travail (appelé aussi répertoire courant) (par exemple `\Mes documents\R` sous windows ou `/Documents/R` sous Mac OS) dans lequel les fichiers de données (`exGraph.dat` ; `auteurDoc.txt` ; `dataQuest.txt` ; `dataRT.txt` ; `dataRT3.txt`)[2] et de fonctions supplémentaires (IPErad ou IPErad.r) seront copiés. Une fois que **R** a été lancé, sélectionnez le répertoire de travail.

### Manipulations de base

Ces manipulations permettent d'effectuer des traitements simples de données issues d'un questionnaire.

*Charger les données et rendre les variables accessibles par leur nom*

```
> quest <- read.table("dataQuest.txt")
> attach(quest)
```

[1] http://www2.unine.ch/ipe/home/liens/utilitaires_statistiques.

[2] Les fichiers d'extension « .dat » contiennent plusieurs tables en format **R** (chargement à l'aide de la fonction load). Les fichiers d'extension « .txt » contiennent une table en format texte (chargement à l'aide de `read.table`).

*Visualiser toutes les données ou en partie*

```
> quest
> quest[1:10,]
> quest[1:10,1:3]
```

*Créer et visualiser une table*

```
> table(age,op1)
> matable <- table(sexe,op1)
> matable
> pretty.table(matable)
```

*Appliquer un test*

```
> chisq.test(matable)
```

**Utiliser la « bibliothèque » IPErad**

Copier dans le répertoire de travail IPErad.r (IPErad). Une fois que **R** a été lancé, sélectionnez le répertoire de travail et exécutez la requête :

```
> source("IPErad")
```

Cette bibliothèque permet d'effectuer les tests présentés dans cet ouvrage. Par exemple pour effectuer le test de Jonkheere (S)

```
> S.test(matable)
```

Les fonctions principales seront présentées systématiquement.

**Diagrammes divers**

*Représentation triangulaire*

* `init.RT()` : prépare une zone graphique (anciennement `init.graph()`)

* `affiche.RT(df,ordre=1,pct=T)` : effectue l'affichage des points à partir d'un « dataframe » construit selon la structure :

- Entête : `"var1" "var2" "var3"`

- Chaque ligne : `"id" v1 v2 v3`

- `pct=T` (vrai par défaut) indique que les données sont déjà des pourcentages (compris entre 0 et 1). Dans ce cas `v3=1-(v1+v2)` peut être omise dans les données.

Cette fonction utilise : `pos.x(ksi,eta)` ; `marque(ksi,eta,c="x",col="red")`

**Exemple**

```
> donnees <- read.table("dataRT.txt")
> init.graph()
> affiche.RT(donnees)
```

* `marque(ksi,eta,c="x",col="red")` : permet l'affichage d'un seul point donné par deux de ses coordonnées « triangulaires », son nom et sa couleur.

*Représentation tétraédrique*

Nécessite la librairie `rgl` (installation : `install.packages("rgl")` puis `library(rgl)`).

* `init.RT3d()` : prépare une zone graphique.

* `affiche.RT3d(df,ordre=1,pct=T)` : effectue l'affichage des points à partir d'un « dataframe » construit selon la structure :

- Entête : `"var1" "var2" "var3" "var4"`
- Chaque ligne : `"id" v1 v2 v3 v4`
- `pct=T` (défaut) indique que les données sont déjà des pourcentages (compris entre 0 et 1). Dans ce cas `v4`=1–(`v1+v2+v3`) peut être omise.

Cette fonction utilise : `pos.3d(a,b,c,d)` ; `marque.3d(a,b,c,d,lab="x",col="red")`

**Exemple**

```
> donnees <- read.table("dataRT3d.txt")
> init.RT3d()
> affiche.RT3d(donnees)
```

*Diagramme de profils*

```
* radar(tab, labels= dimnames(tab)[2],
glabels=dimnames(tab)[1],main=NULL,couleur=
c("red", "blue","green","violet","orange",
 "maroon"),relative=FALSE,rayon=TRUE)
```

```
* profil(tab, labels= dimnames(tab)[2],
glabels=dimnames(tab)[1],main=NULL,couleur=
c("red","blue","green","violet","orange",
"maroon »), relative=FALSE,rayon=TRUE,
radar=TRUE)
```

: affiche un diagramme de profil, chaque ligne de la matrice ou du dataframe tab correspond à un profil.

- `labels` : les noms des variables.
- `glabels` : les noms des observations (les groupes à comparer).
- `main` : titre principal.

- `couleur` : les couleurs des lignes de profil.
- `relative` : `FALSE` le cercle extérieur représente 100% ; `TRUE` le cercle extérieur est adapté au plus grand pourcentage observé
- `radar` : si `TRUE` en format « radar » ou « toile d'araignée »
- `rayon` : si `TRUE`, dessine les rayon de la toile

**Exemple**

```
> dfmm
  X1 X2 X3
1  1  2  3
2  4  5  6
> profil(dfmm,relative=T,main="Essai")
> profil(dfmm,relative=T,labels=c("A","B","C"),
glabels=c("G1","G2"),main="Essai",rayon=F)
```

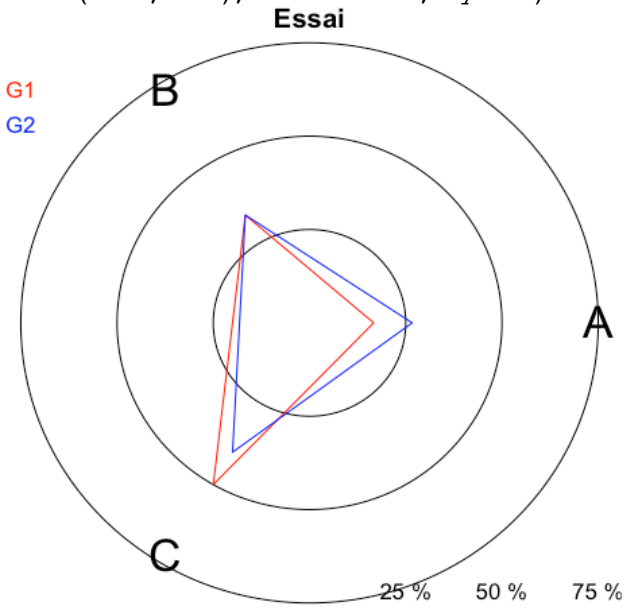



**fig III.18.1**. Profils en format « radar »

## Utilitaires

* `cree.table(ligne,colonne)` permet de créer une table (figure III.18.2).

**Exemple**

```
> tab <- cree.table(2,3)
> tab
```

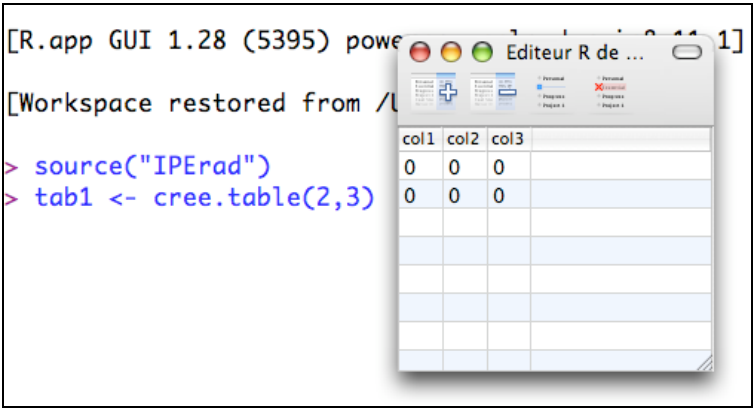



**fig III.18.2**. Edition d'une table

Pour créer une table avec des noms de lignes et/ou colonnes, on introduit les noms à la place des valeurs, par exemple : `cree.table(c("G1","G2"), c("A","B","C"))` ou de façon mixe : `cree.table(c("G1","G2"), 3)`.

Pour éditer un tableau déjà créé, on utilisera `edit(tab)`.

* `pretty.table(tab,pct="row")` affiche une table avec les pourcentages par lignes (défaut), par colonnes (`pct="col"`) ou par rapport au total (`pct="total"`).

* `multinomial(n,decomp)` avec `decomp = c(n1,n2,…nk)` de somme égale à n (nk peut être omis). Cette fonction retourne la valeur : $n!/\prod n_i!$

* `as.boite(tab,plabel, rlabel, clabel)` fabrique une « boîte » à partir de la matrice `tab`. Une boîte est une structure tridimensionnelle constituée d'un ensemble de tables de contingence, chacune étant une « plaque » dans la boîte. Cette structure est utilisée dans les plans factoriels.

`plabel`: étiquette des « plaques » ; `rlabel` : étiquettes des lignes ; `clabel` : étiquettes des colonnes.

**Exemple**

```
> tab26.K <-as.boite(tab26.K,c("S","M","I"),
c("F","D"),c("NC","I","C"))
```

De façon interne, une boîte est un objet qui sépare les informations de structures (nombre d'éléments, étiquettes) des données.

* `aff.boite(boite)` renvoie les données de la « boîte ».

**Statistique non-paramétrique**

*Test basé sur le S de Kendall*

* `S.kendall(tab)` donne la valeur du S liée à la table `tab`.

* `S.test(tab)` effectue un test basé sur le S. Elle utilise : `S.kendall(tab)` ; `prod.S(v,w)` ; `prod.S1(v,w)` ; `var.S(tab)` ; `correc.S(tab)`

**Exemple**

```
> tab
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    3    2    0
[3,]    4    0    1

> S.test(tab)
S = -39, corr = 1, Sc = -38
variance = 363.71, ecart-type = 19.071
z = -1.99, p-value = 0.023

> S.test(table(age,op2))
S = -110, corr = 1, Sc = -109
variance = 15014.59, ecart-type = 122.5
z = -0.89, p-value = 0.19
```

* `distrib.exacte.S(t,u)` donne la distribution exacte de S avec `t` somme de colonnes sous la forme `c(t1, …,tn)` et `u` somme des lignes.

Cette fonction utilise : `distrib0.S(ltab)` ; `multiplicite(tab)` ; `multinomial(n,decomp)` ; `cree.pattern(t,u)` ; `regroupe.tab(tab)`.

* `distrib.exacte.S(tab)` donne la distribution exacte de S pour les tableaux qui ont mêmes totaux marginaux que `tab`..

**Exemple**

```
> distrib.exacte.S(c(3,3,2),c(3,2,3))

     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7] [,8]
[1,] -19   -14   -11   -10    -9    -8    -7   -6
[2,]   3    24     9     18    10    18    36   18

     [,9]  [,10] [,11] [,12] [,13] [,14] [,15]
[1,]  -4    -3    -1     0     1     3     4
[2,]  42    45     3   108     3    45    42

     [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
[1,]   6     7     8     9    10    11    14    19
[2,]  18    36    18    10    18     9    24     3
```

* `S.ftest(boite,plaque2=null)` effectue un test pour une plan factoriel (2 x n). Avec le paramètre `boite` (liste de deux matrices) ou les deux matrices.

Exemple avec les données de la table II.24 (du chapitre II.6)

```
> S.ftest(tab24.fi,tab24.ga)

Analyse de l'interaction avec delta
***********************************
D1 =  -0.292  D2 =  0.00833
delta =  0.301  v =  0.0822
z =  1.05 p =  0.147

> S.ftest(tab24.D,tab24.F)

Analyse de l'interaction avec delta
***********************************
D1 =  -0.385  D2 =  -0.1
delta =  0.285  v =  0.0829
z =  0.989 p =  0.161
```

*Test basé sur le K de Kruskal*

* `K.kruskal(tab)` retourne le vecteur contenant nombre de lignes, nombre de colonnes, valeur de K et facteur correctif associés à `tab`.

* `K.test(tab)` effectue le test de Kruskal-Wallis. Elle utilise : `K.kruskal(tab)` qui retourne `c(nc,nl,K,Tc)`

* `K.ftest(boite,np=NULL,nl=NULL,nc=ncol(boite))` effectue un test pour une plan factoriel (analyse de variance non paramétrique). Avec le paramètre `boite` , une boîte ou une matrice. Dans ce dernier cas, il faut suppléer les paramètres `np`, `nl`, `nc` (respectivement nombre de plaques, de lignes et de colonnes ou listes d'étiquettes).

**Exemple**

Les données sont reprises du chapitre II.8.

```
> K.ftest(tab26.K)
Analyse de variance non parametrique
************************************
T = 0.88  (facteur correctif)
Effet total
***********
Ktotal = 25.13  ( 22.00 ), df = 5, p-value =
0.00013
```

```
Effet  S M I
*************
Kp = 18.94  ( 16.58 ), df = 2, p-value = 7.7e-05
Effet  F D
************
Kr = 1.447  ( 1.267 ), df = 1, p-value = 0.23
Effet d'interaction
******************
Kint = 4.739 , df = 2, p-value = 0.094
```

*Test basé sur le L de Meddis*

* `L.meddis(tab,lambdas)` retourne le vecteur contenant *L*, moyenne, variance et facteur correctif associés à `tab`.

**Exemple**

```
> L.meddis(tab21.L,c(1,2,3))
[1] 138 132  88   1
```

* `L.test(tab,lambdas)` effectue un test basé sur le coefficient *L*.

**Exemple**

```
> L.test(tab21.L,c(1,2,3))
T =  1  L =  138 , mL =  132 ,vL =  88
sc =  9.3808 , z =  0.64 , p-value =  0.26
```

La même fonction permet de traiter les plans factoriels en jouant sur les coefficients lambda.

**Exemple**

```
> L.test(tab25.L,c(1,-1,1,-1))
L =  -103 , mL =  -183 , vL =  18117 ,
sc  =  125.7137 , z =  0.6363668 , p-value =
0.26

> L.test(tab25.L,c(1,1,-1,-1))
L =  331 , mL =  122 , vL =  18218.67
sc  =  126.0659 , z =  1.657863 , p-value =
0.049

> L.test(tab25.L,c(1,-1,-1,1) )
L =  -446 , mL =  -305 , vL =  17791.67
sc  =  124.5798 , z =  1.131805 , p-value =  0.13
```

*Test basé sur le chi2*

* `Chi2.test(tab)` effectue le test du chi2. Elle utilise `Chi2(tab)` qui retourne `c(nl, nc, chi2)`. La fonction **R** de base, chisq.test, est plus complète.

*Test pour échantillons appariés*

* `Z.test(tab,av,ap,a,pct=TRUE)` effectue un test sur une progression de `av` à `ap` avec les probabilités de progressions données par `tab` en pourcent ou non. Elle passe par une approximation normale et utilise `distrib.Z(tab,av,a,pct)`.

**Exemple**

Les données sont reprises du chapitre II.4.

La probabilité d'évolution « naturelle » est donnée dans deux cas :

```
> prog1
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    1    1    0
> prog2
     [,1]  [,2]  [,3]
[1,]  0.5 0.333 0.1667
[2,]  0.5 0.500 0.0000
```

- `prog1` indique que les progressions de 0, 1 ou 2 pour un individu de niveau 1 ont le même poids (1e ligne). Il en va de même pour un individu de niveau 2 (mais il ne peut progresser de 2 niveaux).

- `prog2` indique les probabilités de progressions. Dans cet exemple, les probabilités pour un individu de niveau 1 de ne pas progresser, de progresser de 1 niveau ou de 2 niveaux sont respectivement 0.5, 0.333 et 0.1667.

Probabilité de l'évaluation constatée :

```
> Z.test(prog1,c(7,11,4),c(1,7,14),2,pct=F)
moy:  12.5 ; ecart-type:  2.723356
diff:  16 ; z:  1.285179 ; p-value:  0.09936485

> Z.test(prog2,c(7,11,4),c(1,7,14),2)
moy:  10.16667 ; ecart-type:  2.576604
diff:  16 ; z:  2.263962 ; p-value:  0.01178823
```

* `distrib.exacte.Z(v,a)` donne la distribution exacte des progressions possibles d'empan `a` à partir de la répartition `v`. Elle utilise : `distrib0.Z(tab,a)` (la ligne i de `tab` contient la place du sujet i) ; `distrib1.Z(tab,a)` ; `enumLM(v)` (passage d'une énumération à une forme matricielle) ; `somme.rang(tab)`.

*Calcul des effets*

* `effet(tab,coef,dim=1)` donne la valeur de l'effet calculée pour les coefficients (coef) "`gamma`", "`delta`" (ligne, colonne ou symétrique pour `dim` valant respectivement 1, 2 ou 3), "`Phi`", "`C`", "`tau.b`", "`tau.c`".

Pour chaque coefficient, `coef`, la fonction utilise la fonction `IPE.<coef>`.

*Fonctions statistiques diverses*

* `cree.pattern(t,u)` : crée tous les tableaux possibles avec `u` étant somme des colonnes et `t` la somme des lignes.

* `cree.pattern2(t,m)` : crée tous les tableaux possibles de `m` colonnes avec `t` valant la somme des lignes.

* `moy.tab(tab)` : crée le tableau avec les fréquences proportionnelles aux totaux lignes et colonnes de `tab` (tableau de contingence).

* `valeurs.chi()` : utiliser pour des tests.

* `cronbach(tab)` : calcul du coefficient de Cronbach.

* `pct.table(tab,pct="row")` : calcul des pourcentages de `tab` en ligne, colonne ou selon total.

**Calcul de l'entropie**

`H(v,relative=F)` : entropie de Shannon, avec `v = c(`$n_1$`, `$n_2$`, ..., `$n_K$`)`. La valeur relative est obtenue en faisant le rapport à la valeur maximale.

`HP(v,relative=F)` : entropie selon la formule de Shannon avec utilisation du logarithme naturelle (valeur de l'entropie de Shannon divisée par ln 2). La valeur relative est obtenue en faisant le rapport à la valeur maximale.

`SB(v)` : entropie calculée selon la formule de Boltzmann.

`HH(v,exact=T)` : entropie selon la qualité calculée de manière exacte ou approchée.

`HV(x,v)` : entropie sur l'utilité avec `x` vecteur des utilités et `v` vecteur contenant les effectifs des groupes.

**Représentation de graphes**

Nécessite la librairie `igraph` et accessoirement la librairie `tcltk` .

* `core(adj,arr=0)` : donne la cardinalité du « core » lié à chaque sommets du graphe qui lui est donné par la matrice d'adjacence `adj` . (utilise `mcore`).

* `fermeture(adj,thres=0,arr=0)` : fermeture transitive du graphe selon la technique proposée par Pochon & Favre (2009).

* `centralite(adj,thres=0,ferm=F,arr=0)` : donne les coefficients de centralité entrante et sortante pour chaque sommet du graphe de matrice d'adjacence `adj`. `ferm` indique si la fermeture a déjà été effectuée (par défaut une fermeture transitive est effectuée). Les autres paramètres sont destinés à la fermeture.

* `gr <- graph.adjacency(adj,…)` : crée un graphe à partir d'une matrice d'adjacence (`igraph`) que l'on peut afficher avec `plot`.

* `tklplot(gr)` : affiche un graphe dans un format qu'il est permis de modifier (librairie `tcltk` après avoir précisé le *device* `X11()` )

**Exemple**

```
> load("exGraph.dat")
> library(igraph)
> gr20 <- graph.adjacency(mat20,mode="dir")
> plot(gr20)

> centralite(mat20,ferm=F)
> centralite(mat20)
> core(mat20)
```

## Table des matières