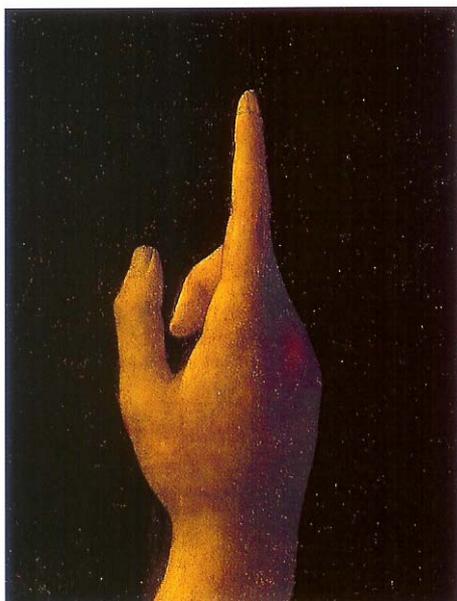


La Statistique Pratique

ou

La statistique à la portée de tous



**André CAMLONG
&
Christine CAMLONG-VIOT**

2006

Ouvrage cordialement dédié aux

« Amis du STABLEX »

et affectueusement

à Léo et à Florian

Photo couverture :
Léonard de Vinci
Saint Jean-Baptiste, 1513-16.
Paris, Musée du Louvre

Entièrement conçu et réalisé sur PC
par
André CAMLONG
&
Christine CAMLONG-VIOT

AVANT-PROPOS

par

André CAMLONG

Professeur Titulaire à l'Université de Toulouse II

&

Christine CAMLONG-VIOT

Maître de Conférences à l'Université de Paris XI

La statistique à la portée de tous est un ouvrage destiné à un large public, dès lors qu'il met « *la pratique statistique à la portée de tous* ». C'est un ouvrage essentiellement pratique, même s'il est rigoureux et technique. C'est justement ce côté technique qui effraie bien souvent l'utilisateur que nous avons voulu surmonter pour mettre la pratique statistique à la portée de tous, et notamment aider à la compréhension de *SATBLEX* et en faciliter l'utilisation.

1. *La statistique à la portée de tous* est d'abord un outil de description et d'analyse, c'est ensuite un outil de décision et d'aide à l'interprétation, c'est enfin un outil de mesure et de contrôle des données, qui fait appel conjointement à la puissance de l'outil informatique et de l'outil mathématique et statistique pour mener à bien une analyse rigoureuse des données statistiques, une « analyse scientifique » *stricto sensu*.

2. *La statistique à la portée de tous* n'est ni un cours de statistique ni un ouvrage pédagogique, à proprement parler. On suppose que le lecteur est déjà familiarisé avec l'informatique et avec les statistiques. En cas de doute ou de nécessité, l'utilisateur pourra se reporter utilement à des livres spécialisés. Du point de vue technique, on suppose également que l'utilisateur est déjà familiarisé avec Excel, et qu'il manipule ce logiciel avec aisance. Néanmoins, pour faciliter le travail de tout utilisateur, nous avons mis à sa disposition une Macro qui effectue automatiquement toutes les tâches, notamment les calculs les plus ardues.

3. *La statistique à la portée de tous* ouvre les voies de l'analyse multidimensionnelle des données suivant des méthodes exploratoires, paramétriques, objectives et inductives, faisant de la statistique un outil d'analyse, de description et d'aide à l'interprétation.

4. *La statistique à la portée de tous* se compose de 7 chapitres indépendants, mais qui s'enchaînent, abordant le traitement complet de données statistiques concrètes :

– **Chap. 1. Tables de Contingence et Densités Métriques**, c'est l'étude pratique et la pratique d'une analyse appliquée au traitement d'une population de 10 000 étudiants concernant le choix des disciplines universitaires en fonction des origines sociales.

– **Chap. 2. Ajustement, Corrélation et Analyse en Composantes Principales (ACP)**, c'est l'étude de l'Ajustement des variables, de la Corrélation et à l'Analyse en Composantes Principales pratiquée sur cette même base de données.

– **Chap. 3. La Régression linéaire**, c'est la pratique de l'analyse linéaire ou factorielle discriminante (AFD).

- **Chap. 4. Images et Mesures d’Inertie ou Procédures d’Analyse Discriminante**, c’est l’étude des images de synthèse décrivant l’inertie vectorielle entamée à la fin du chapitre précédent, concernant les procédures d’Analyse Discriminante qui sont au cœur de l’Analyse Statistique.
- **Chap. 5. Traitement d’un corpus complexe ou multidimensionnel**, c’est l’étude d’un corpus dont la dimension linéaire est complexe, parce que formée de vecteurs simples ou de vecteurs composites, comme dans le cas du corpus analysé, tiré des *Huit contes en prose* de Perrault..
- **Chap. 6. Mode d’emploi de la statistique analytique**, c’est l’analyse exploratoire d’une base de données complexes et multidimensionnelle, effectuée en tournant les pages de la Macro proposée, appliquée à l’étude d’une base de données concernant l’infarctus du myocarde.
- **Chap. 7. La Régression linéaire appliquée à l’étude d’un échantillon**. Ce dernier chapitre tente de familiariser l’utilisateur avec la détermination des intervalles de confiance pour une meilleure utilisation d’Excel.
- **Chap. 8. Formulaire**. Formules de base, décomposition de la variance, distance dt , distances quadratiques, inertie totale, vecteurs isotropes, corrélations et ajustements, résidus, spectres de décomposition.
- **Chap. 9. Régression linéaire simple**. Étude de cas : *Le Petit Chaperon rouge* de Perrault par le biais de l’ACP (Analyse en Composantes Principales) et de l’AFD (Analyse Factorielle Discriminante).
- **Chap. 10. Métrique R et ACP**. Descriptions des formules et applications.
- **Chap. 11. Formulaire d’Excel** concernant les TESTS STATISTIQUES. Application et justification des formules.

5. *La statistique à la portée de tous* a la prétention de mettre la statistique à la portée de tous en montrant, par la pratique, comment la puissance de l’outil informatique et la puissance de l’outil mathématique et statistique sont au service du plus grand nombre pour traiter méthodiquement et rigoureusement tout corpus de données statistiques élaboré suivant les règles de l’art.

Nous avons en outre la faiblesse de croire que le pari est gagné d’avance, grâce à la Macro que nous proposons. Comme la pratique vaut mieux que les beaux discours du monde, nous invitons l’ utilisateur à ouvrir cet ouvrage en même temps que son ordinateur et de lancer la Macro pour traiter le corpus de ses données.

Tous les moyens sont ici réunis pour que tout un chacun puisse maîtriser l’outil statistique en vue d’une parfaite analyse des données d’où il tirera des conclusions parfaitement contrôlées, toujours vérifiées et toujours vérifiables.

Toulouse, mai 2006
Les Auteurs

AC & CC

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

1

Tables de Contingence et Densités Métriques

par
André CAMLONG
Christine CAMLONG-VIOT

La statistique à la portée de tous ceux qui veulent s'adonner à la pratique de l'analyse « scientifique » des phénomènes observés concernant une population donnée consignés dans des tables de contingence à partir des relevés méthodiques et exhaustifs.

Que faire de ces données ? Comment les analyser ? Avec quels outils ? Comment les interpréter ? Quelle en est la signification ? Quelle en est la portée ?

Autant de questions qu'on se pose légitimement à partir des relevés ou des recensements objectifs concernant des « populations » de tous ordres et de toute nature.

Comment y répondre ? C'est justement l'objectif que nous nous sommes fixés d'atteindre et que nous proposons ci-après en même temps qu'une Macro pour y faire face.

Rappelons que la statistique n'est pas une fin en soi, mais un moyen d'analyse et d'aide à l'interprétation. Elle ne se prononce jamais sur l'essence des phénomènes, mais sur la qualité des phénomènes eux-mêmes qu'elle mesure, compare et permet d'interpréter avec force calculs et graphiques.

Pour ce faire, la statistique dispose aujourd'hui de la puissance du calcul informatique qui valorise à son tour la puissance du calcul mathématique dans ses trois composantes que sont le calcul algébrique qui permet de mesurer, de comparer et d'intégrer les données dans un tout cohérent et complet (l'algèbre étant, étymologiquement, la réduction de la fraction à l'intégralité) ; le calcul arithmétique qui permet d'évaluer, de déterminer et de contrôler les calculs d'analyse ; et la représentation géométrique qui permet de visualiser, de mémoriser et de raisonner sur la qualité des phénomènes à partir d'images réelles et sensibles.

Pour ce faire, la statistique dispose d'Excel, un outil informatique « tout puissant », assorti de fonctions prédéfinies et des Macros que l'on peut soi-même adapter à son travail.

Voilà pourquoi, au terme des exposés qui vont suivre, nous proposons au lecteur une Macro et un logiciel STABLEX qui lui faciliteront grandement l'accès à l'informatique et à la statistique pour s'adonner à l'analyse de tout type de corpus.

Cette approche va se faire en quatre temps et trois mouvements, par le biais des 6 chapitres intitulés : 1) les Tables de Contingence et les Calculs des Densités métriques ; 2) les Corrélations et l'Analyse en Composantes Principales (ACP) ; 3) la Régression et les Profils de Densité ; 4) les Images et les Points d'Inertie ; 5) la Discrimination et la Lemmatisation ; 6) le Mode d'emploi de la statistique analytique.

Pour être plus concrets et plus pratiques encore, nous prenons comme support pédagogique les données utilisées par G. SAPORTA dans *Probabilités, Analyse des Données et Statistique*, publié aux Editions TECHNIP, Paris, 1990, à la page 151, tirées des « *Données Sociales* », 3° édition, INSEE, 1978. Nous prendrons ensuite des données plus complexes. Le tout étant de ne pas perdre de vue qu'il s'agit de mettre *la statistique à la portée de tous*.

*

* *

1. La Base des Données

Il s'agit d'un « échantillon » concernant les origines sociales de 10 000 étudiants en 1975-1976 répartis en 9 classes sociales et inscrits dans 8 disciplines différentes.

Discipline / CI	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
D1 : Droit	80	6	168	470	236	145	166	16	305
D2 : Sciences Eco	36	2	74	191	99	52	64	6	115
D3 : Lettres	134	15	312	806	493	281	401	27	624
D4 : Sciences	99	6	137	400	264	133	193	11	247
D5 : Médecine-Dentaire	65	4	208	876	281	135	127	8	301
D6 : Pharmacie	28	1	53	164	56	30	23	2	47
D7 : Pluridisciplinaire	11	1	21	45	36	20	28	2	42
D8 : IUT	58	4	62	79	87	54	129	8	90

Les classes sociales sont portées dans les colonnes et numérotées de 1 à 9 suivant la nomenclature CI 1, CI 2, ..., CI 9 :

CI 1	Exploitants agricoles
CI 2	Salariés agricoles
CI 3	Patrons
CI 4	Professions libérales et Cadres supérieurs
CI 5	Cadres moyens
CI 6	Employés
CI 7	Ouvriers
CI 8	Personnel de service
CI 9	Autres

Et les disciplines sont portées dans les lignes suivant la nomenclature D1, D2, ..., D8 :

D1	Droit
D2	Sciences Eco
D3	Lettres
D4	Sciences
D5	Médecine-Dentaire
D6	Pharmacie
D7	Pluridisciplinaire
D8	IUT

2. La Table de Distribution des Fréquences (la TDF)

La Table de Distribution des Fréquences (TDF) reprend les valeurs de la base des données en y ajoutant les valeurs marginales qui vont servir aux calculs de probabilités et aux calculs des écarts centrés réduits. Elle fait abstraction des dénominations adjacentes, comme, dans le cas présent, celles des classes sociales et des disciplines.

D	E	F	G	H	I	J	K	L	M
Total	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
1592	80	6	168	470	236	145	166	16	305
639	36	2	74	191	99	52	64	6	115
3093	134	15	312	806	493	281	401	27	624
1490	99	6	137	400	264	133	193	11	247
2005	65	4	208	876	281	135	127	8	301
404	28	1	53	164	56	30	23	2	47
206	11	1	21	45	36	20	28	2	42
571	58	4	62	79	87	54	129	8	90
10000	511	39	1035	3031	1552	850	1131	80	1771

Les valeurs marginales horizontales (des colonnes) permettent, dans un premier temps, de calculer les probabilités « p » et les probabilités contraires « $q = 1 - p$ » de chaque variable aléatoire, telles que présentées dans les trois lignes suivantes :

	D	E	F	G	H	I	J	K	L	M
1	10000	511	39	1035	3031	1552	850	1131	80	1771
2	p	0,0511	0,0039	0,1035	0,3031	0,1552	0,085	0,1131	0,008	0,1771
3	q	0,9489	0,9961	0,8965	0,6969	0,8448	0,915	0,8869	0,992	0,8229

Les reports dans la colonne D sont ceux de la Macro qui automatise les calculs.

Ces valeurs marginales (des lignes) permettent, dans un deuxième temps, de calculer les valeurs centrées et réduites correspondant aux valeurs de la TDF et de les consigner dans une nouvelle table de contingence, la TDR (Table des Ecartés Réduits), de telle sorte qu'il y a un lien très étroit entre les données brutes de la TDF et les densités métriques de la TDR.

3. La Table des Écarts-Réduits (la TDR)

En fonction des probabilités « p » et « q » (par colonne), il est aisé de calculer les écarts centrés et réduits pour chacune des cases de la TDF, en appliquant la formule classique de calcul du « z » (la valeur centrée et réduite de la densité métrique) :

$$z = \frac{x - \bar{x}}{\sigma_x} \text{ ou } z = \frac{x - \bar{x}}{\sqrt{npq}}$$

Comment procéder du point de vue technique ?

La programmation se fait d'abord dans la première case de la table, celle qui correspond ici aux 80 occurrences de C1 1 (colonne E) pour la première discipline D1, le *Droit* (ligne 6), par rapport au total de la ligne (valeur marginale de la colonne D) :

$z = (E6 - \$D6 * E\$2) / \text{RACINE}(\$D6 * E\$2 * E\$3)$
--

Prendre soin de distinguer les valeurs absolues et les valeurs relatives pour pouvoir étendre le calcul à l'ensemble des cases de la table.

Dans la première case, $z = -0,154$. Cette formule, étendue à la plage correspondante de la TDF, donne automatiquement les valeurs « z » des écarts centrés et réduits de la TDR.

On calcule alors le khi2 de Fisher pour déterminer la « normalité » de la distribution.

Voyons une copie des résultats fournis par la Macro :

C	D	E	F	G	H	I	J	K	L	M
3,767	Total	4,664	0,112	1,531	-2,643	0,008	0,129	1,710	0,684	-2,427
0,471	Moy	0,583	0,014	0,191	-0,330	0,001	0,016	0,214	0,086	-0,303
0,631	Khi2	0,340	0,000	0,037	0,109	0,000	0,000	0,046	0,007	0,092
1										
Ecart :	Moy	Max	Min		Borne inf	Borne sup				
	0,471	13,036	-8,566		-2,000	2,000				
					11	9	27,78%			
Rang	Moy	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
1	0,085	-0,154	-0,084	0,266	-0,684	-0,767	0,870	-1,112	0,918	1,514
2	0,031	0,601	-0,312	1,021	-0,231	-0,019	-0,328	-1,033	0,394	0,190
3	0,225	-1,964	0,847	-0,480	-5,144	0,644	1,167	2,906	0,455	3,590
4	0,213	2,690	0,079	-1,464	-2,910	2,343	0,590	2,002	-0,268	-1,145
5	-1,001	-3,799	-1,369	0,035	13,036	-1,861	-2,837	-7,035	-2,016	-3,164
6	-0,180	1,662	-0,459	1,827	4,498	-0,921	-0,774	-3,565	-0,688	-3,199
7	0,152	0,150	0,220	-0,073	-2,644	0,775	0,622	1,034	0,275	1,007
8	0,893	5,478	1,191	0,399	-8,566	-0,187	0,820	8,512	1,612	-1,219

Dans les trois premières lignes figurent :

- 1) le total des valeurs de « z » des 9 variables ;
- 2) la moyenne des « z » pour chaque variable ;
- 3) le khi2 (carré de la moyenne).

Avec une valeur de 0,631 en C3, le khi2 dit que la distribution, à 9 ddl (degrés de liberté), est normale avec une probabilité de 100 % (0,999917719 arrondi à 1). Et donc que les résultats de l'analyse statistique seront fiables à 100 %. (Nous reviendrons plus tard sur la différence entre le nombre de degrés de liberté d'un échantillon ($n - 1$) et (n), celui d'un corpus intégral, car la notion d'échantillon est souvent galvaudée).

Les deux lignes suivantes donnent, en rouge, les valeurs statistiques de la Moyenne, du Maximum et du Minimum.

Les deux boutons qui s'y trouvent permettent de faire varier les intervalles d'estimation des « z » pour détacher en rouge les valeurs positives et en bleu les valeurs négatives, et faciliter par là même la lecture de la table de contingence.

4. Lectures croisées de la TDR

Il n'est pas question ici d'analyser intégralement le corpus retenu, mais de poursuivre le but pédagogique d'une utilisation pratique de la statistique.

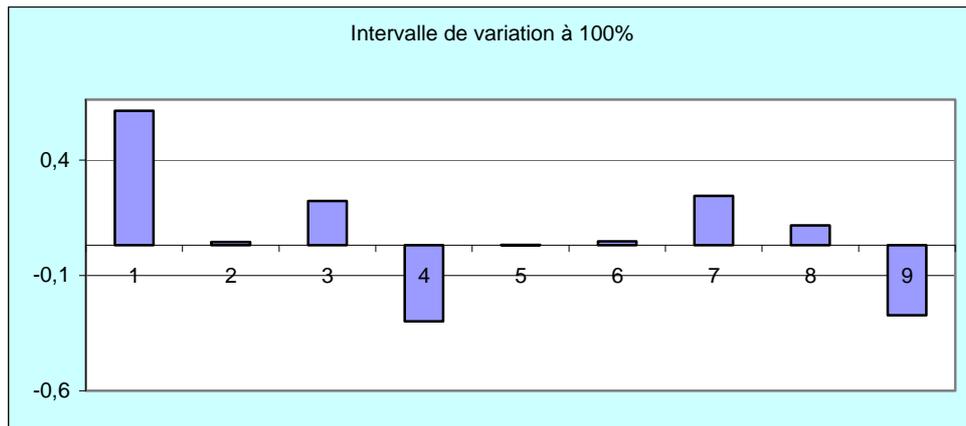
4.1 Normalité de la distribution

La normalité de la distribution est fixée par le test du khi2 de Fisher (qu'on ne confondra pas avec le khi2 de Pearson).

Avec une valeur de 0,631 à 9 ddl, il est aisé de fixer les limites de l'intervalle de variation entre $\pm 0,631$ autour de la moyenne grâce aux échelles graphiques proposées par Excel.

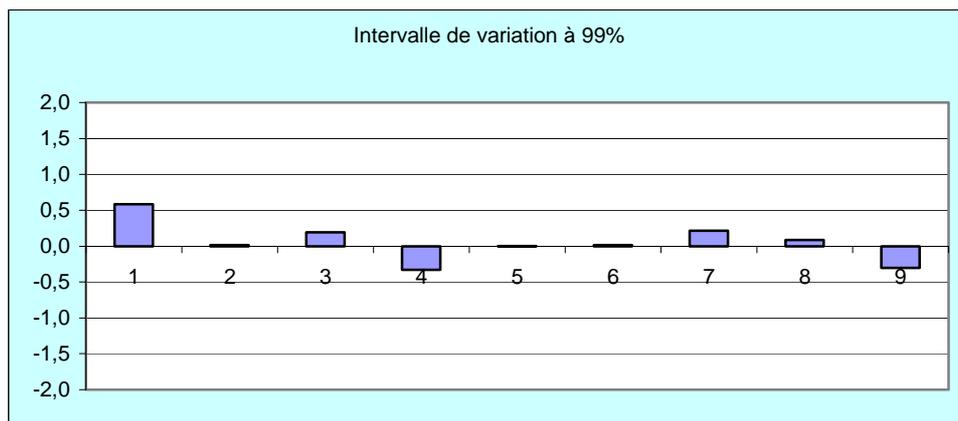
Le graphique sous forme d'histogramme ou de barres permet d'embrasser d'un seul coup d'œil les qualités de la dispersion.

Prenons le cas d'une probabilité à 100% (à 9 ddl), les limites de variation sont dans l'intervalle $\pm 0,631$:

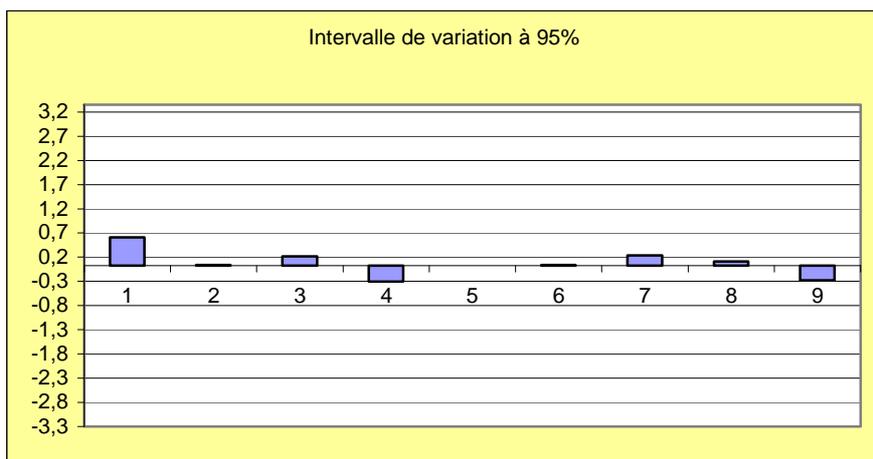


On voit immédiatement quelle est la poussée des variables autour de la moyenne.

Prenons le cas d'une probabilité à 99% (à 9 ddl), les limites de variation sont dans l'intervalle $\pm 2,088$. On voit que la représentation des variables tend à s'écraser :



Prenons le cas d'une probabilité à 95%, les limites de variation sont repoussées à $\pm 3,325$. L'histogramme est aplati, les variables se resserrent autour de la moyenne :



La distribution étant « normale », l'analyse statistique peut se poursuivre, les résultats seront pertinents.

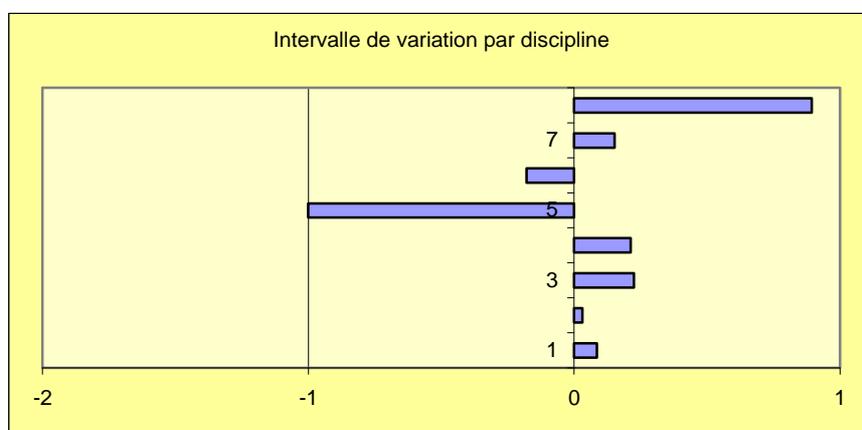
Quelle en est la signification sur le plan horizontal ? Origine des classes sociales ?

Des 9 classes sociales, 2 seulement affichent une densité négative – Cl 4 (professions libérales et cadres supérieurs) et Cl 9 (les Autres). Ces 2 classes n'accordent aucune priorité aux études énoncées. En revanche, les 7 autres classes, avec une densité positive, bien qu'à des degrés différents, manifestent un certain intérêt pour les études. La classe Cl 1 (enfants des exploitants agricoles), c'est celle qui accorde le plus d'intérêt aux études. Les autres classes sont plus nuancées ou moins marquées.

Quelle en est la signification sur le plan vertical ? Choix des disciplines étudiées ?

On fera le même type d'observations pour ce qui est des études poursuivies.

Le khi2 de Fisher, avec une valeur de 1,959 à 8 ddl, et une probabilité à 98,22%, montre que le choix des disciplines se fait de façon « normale », comme il ressort du graphique ci-après :

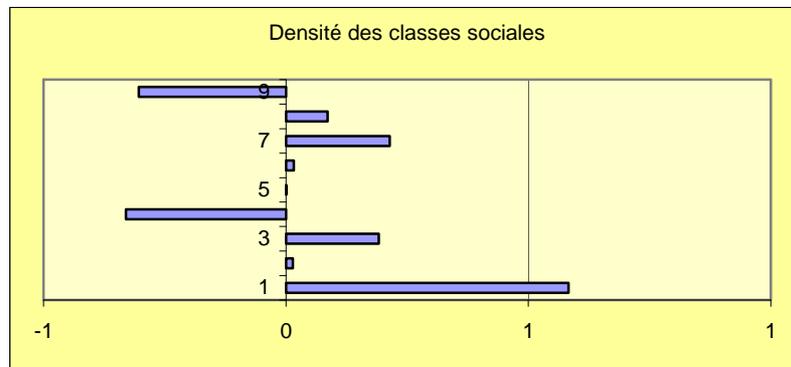


Deux disciplines, D5 (études de Médecine et Dentaire) et D6 (études de Pharmacie), avec des valeurs négatives, ont tendance à repousser les étudiants, alors que les autres disciplines les attirent, notamment les IUT en D8.

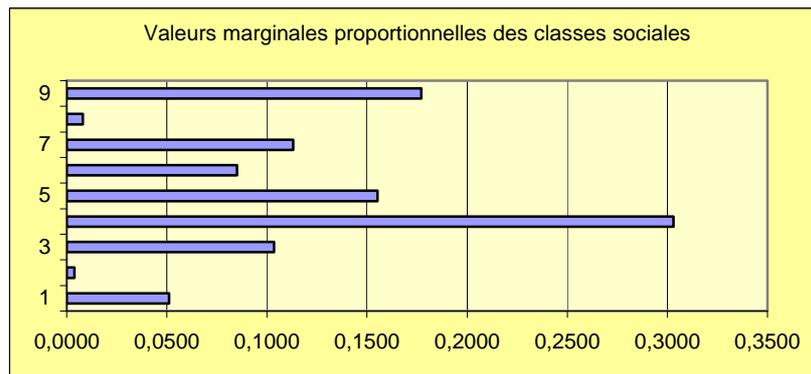
Telle est la signification de la valeur algébrique, résultant d'un calcul d'intégration et de comparaison, alors que les pourcentages ou les proportions n'ont aucune valeur comparative.

Pour s'en convaincre, il suffit de comparer les valeurs algébriques de la TDR aux valeurs correspondantes de la TDF, les unes étant « mesurées » autour de la moyenne et les autres alignées dans le même sens. Tel est le cas, par exemple, de la densité des choix faits par les classes sociales dans la TDR comparée aux rapports proportionnels de ces mêmes valeurs marginales (brutes) de la TDF :

1) la densité des choix de la TDR :



2) les valeurs marginales brutes de la TDF :



Les valeurs brutes de la TDF ne permettent pas de se prononcer sur l'impact des études sur les classes sociales, contrairement aux valeurs « relatives » de la TDR qui accusent les préférences. La classe CI 4 (professions libérales et des cadres supérieurs) arrive en tête des valeurs absolues, mais la TDR montre que malgré l'importance des effectifs, cette classe est amplement déficitaire. Il en va de même de CI 9, « Autres », qui arrive dans l'absolu en deuxième position, mais accuse un déficit de densité visible.

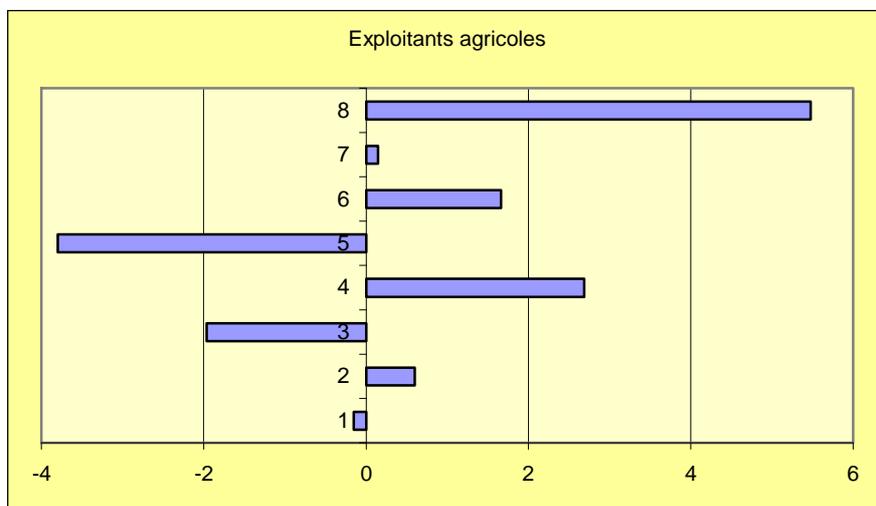
Ce type d'observation est valable aussi bien pour les lignes que pour les colonnes, pour le choix des disciplines aussi bien que pour l'attrait des études exprimé par les classes sociales.

4.2 Lecture verticale : choix des disciplines par les différentes classes sociales

Cl 1 (fils d'exploitant agricole) est attirée par D4 (les Sciences) et surtout D8 (IUT) suivant les valeurs significatives positives, en rouge, mais rejette D5 (les études de médecine et dentaires) et D3 (les Lettres) avec des valeurs significatives négatives, en bleu.

Cl 1	Exploitant agricole
-0,154	D1 : Droit
0,601	D2 : Sciences Eco
-1,964	D3 : Lettres
2,690	D4 : Sciences
-3,799	D5 : Médecine-Dentaire
1,662	D6 : Pharmacie
0,150	D7 : Pluridisciplinaire
5,478	D8 : IUT

Le graphique indique combien les préférences disciplinaires de Cl 1 sont marquées :

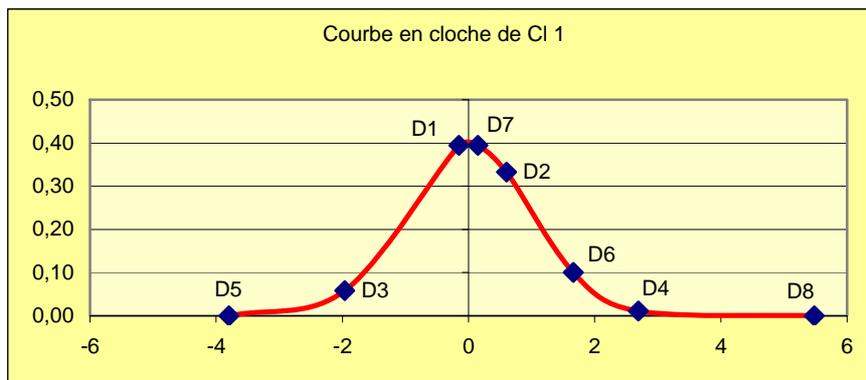


D8 (IUT) et D4 (les sciences) avec des valeurs significativement positives, montre combien le choix est important. À l'inverse, D5 (médecine et dentaire) et D3 (lettres) avec des valeurs significativement négative, montre à l'inverse combien ce type d'étude est rejeté par les fils d'exploitants agricoles.

On pourrait tout aussi bien détacher un ordre préférentiel global :

CI 1	Loi normale	Discipline
-3,799	0,242	D5 : Médecine-Dentaire
-1,964	0,389	D3 : Lettres
-0,154	0,397	D1 : Droit
0,150	0,394	D7 : Pluridisciplinaire
0,601	0,399	D2 : Sciences Eco
1,662	0,393	D6 : Pharmacie
2,690	0,390	D4 : Sciences
5,478	0,268	D8 : IUT

et l'accompagner d'une courbe de Laplace-Gauss en cloche :



Comment réaliser cette courbe ? Il faut :

- 1) ordonner par ordre croissant les valeurs algébriques de CI 1 (colonne) ;
- 2) sur une colonne parallèle calculer les valeurs correspondantes de la « LOI.NORMALE » en prenant pour paramètres les valeurs centrées de « z », avec « espérance 0, écart-type 1 » et la fonction « cumulative » FAUX ;
- 3) sélectionner les deux colonnes et faire le graphe en « nuages de points ».

On analysera de la même façon chacune des lignes et chacune des colonnes de la TDF et de la TDR correspondante. On déterminera ainsi les choix préférentiels de chacune des classes sociales.

4.3 Lecture horizontale : les origines sociales en fonction des disciplines

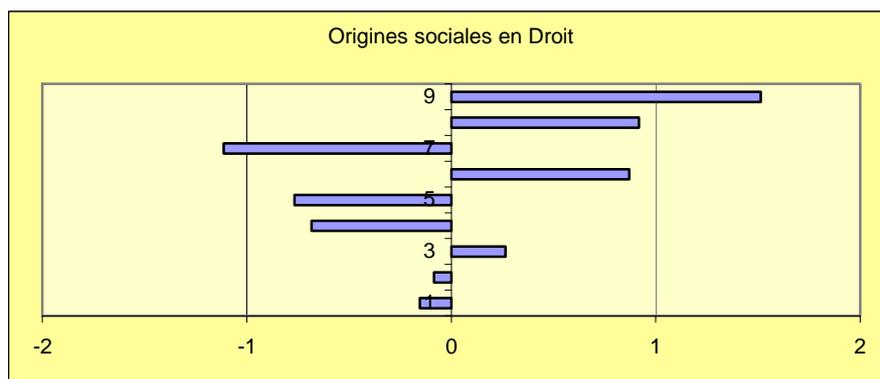
De même, la TDR met en évidence les préférences disciplinaires des classes sociales.

Prenons le cas du Droit (ligne D1) fréquenté par toutes les classes sociales, mais pas au même degré ou avec la même densité :

Droit	Classe	Nature
-0,154	CI 1	Exploitants agricoles
-0,084	CI 2	Salariés agricoles
0,266	CI 3	Patrons

-0,684	CI 4	Professions libérales et Cadres supérieurs
-0,767	CI 5	Cadres moyens
0,870	CI 6	Employés
-1,112	CI 7	Ouvriers
0,918	CI 8	Personnel de service
1,514	CI 9	Autres

Le graphique en barres rend parfaitement compte des qualités de la dispersion :

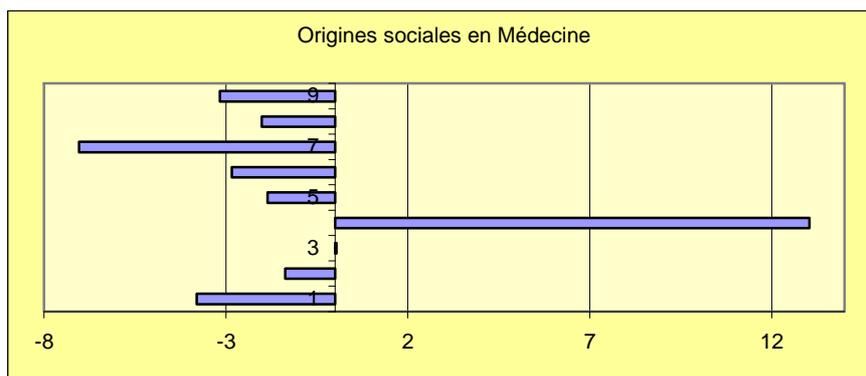


Le Droit (D1) exerce un attrait particulier sur les classes CI 9 « Autres », CI 8 (personnel de service) ou sur CI 6 (employés), et, à l'opposé, une aversion sur les classes CI 7 (ouvriers), CI 5 (cadres moyens) ou CI 4 (professions libérales et cadres supérieurs).

Le cas des « études de médecine et dentaires » est typique. On dirait que ces études sont réservées à une classe, CI 4, « professions libérales et cadres supérieurs » :

Médecine	Classe	Nature
-3,799	CI 1	Exploitants agricoles
-1,369	CI 2	Salariés agricoles
0,035	CI 3	Patrons
13,036	CI 4	Professions libérales et Cadres supérieurs
-1,861	CI 5	Cadres moyens
-2,837	CI 6	Employés
-7,035	CI 7	Ouvriers
-2,016	CI 8	Personnel de service
-3,164	CI 9	Autres

Le graphique en barres est à cet égard éloquent. Avec un $z = +13,036$ on voit qu'il s'agit d'une discipline « réservée » ou « exclusive », comme si elle était réservée à une caste, celle des « professions libérales et des cadres supérieurs » :



On procéderait de la même façon pour analyser les autres lignes et les autres colonnes.

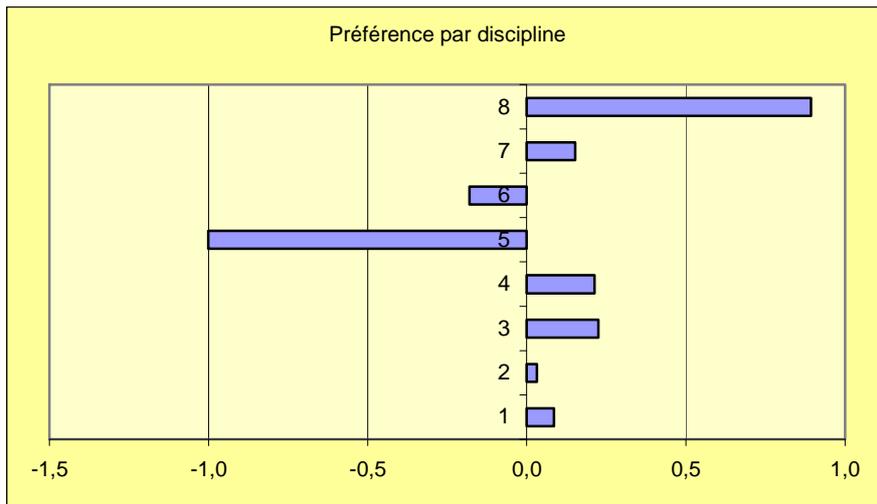
Il est évident que toutes les valeurs algébriques de la TDR ont une haute portée comparative, contrairement aux valeurs absolues de la TDF ou des pourcentages que l'on pourrait en déduire.

5. Profil général

Le profil général est donné par les moyennes des valeurs algébriques marginales de la TDR :

Préférence	Discipline
0,085	D1 : Droit
0,031	D2 : Sciences Eco
0,225	D3 : Lettres
0,213	D4 : Sciences
-1,001	D5 : Médecine. Dentaire
-0,180	D6 : Pharmacie
0,152	D7 : Pluridisciplinaire
0,893	D8 : IUT

Le graphique qui lui est associé permet de visualiser immédiatement les choix des 8 disciplines fait par les 9 classes sociales représentées :

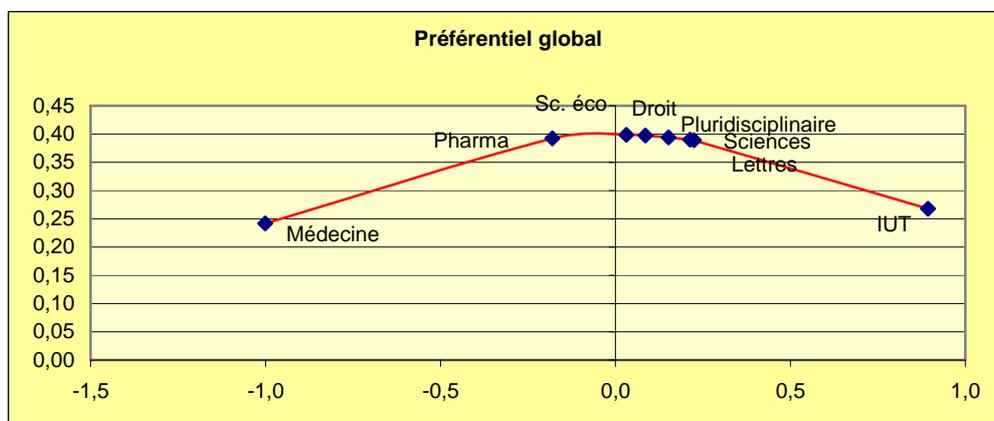


La courbe en cloche de Laplace-Gauss permet de généraliser la vision :

- 1) on fait le tableau préférentiel en rangeant les valeurs algébriques par ordre décroissant et en identifiant les disciplines :

Moyenne	Loi normale	Discipline
-1,001	0,242	D5 : Médecine-Dentaire
-0,180	0,393	D6 : Pharmacie
0,031	0,399	D2 : Sciences Eco
0,085	0,397	D1 : Droit
0,152	0,394	D7 : Pluridisciplinaire
0,213	0,390	D4 : Sciences
0,225	0,389	D3 : Lettres
0,893	0,268	D8 : IUT

- 2) on trace la courbe sous forme de cloche de Laplace-Gauss :



6. Caractéristiques de chaque variable par rapport à l'ensemble de définition

Le calcul algébrique produit par la TDR intègre chaque variable dans l'ensemble de définition (que nous considérerons comme un corpus entier et non comme un échantillon au sens traditionnel du terme).

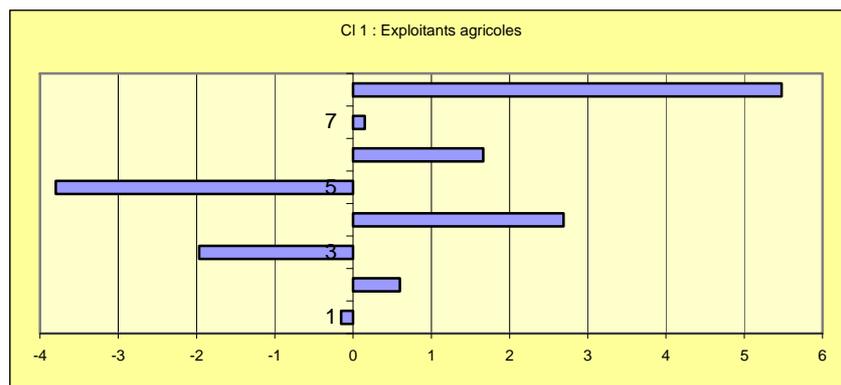
Il ressort que les valeurs algébriques sont immédiatement décryptées et qu'on peut lire la TDR aussi bien horizontalement que verticalement ; aussi bien globalement (valeurs moyennes marginales) que particulièrement, ligne après ligne, colonne après colonne, ou case après case.

Les valeurs algébriques donnent la mesure exacte de la comparaison intégrale : toutes ces mesures sont à la mesure de la dimension quanti-qualitative de la distribution. Il suffit de se reporter aux valeurs limites de signification suivant les seuils de probabilité pour évaluer à sa juste mesure chaque variable, chaque vecteur ou chaque facteur. Et ensuite le représenter par un graphique « qui parle » pour en percevoir les contours et se lancer dans l'analyse des phénomènes observés.

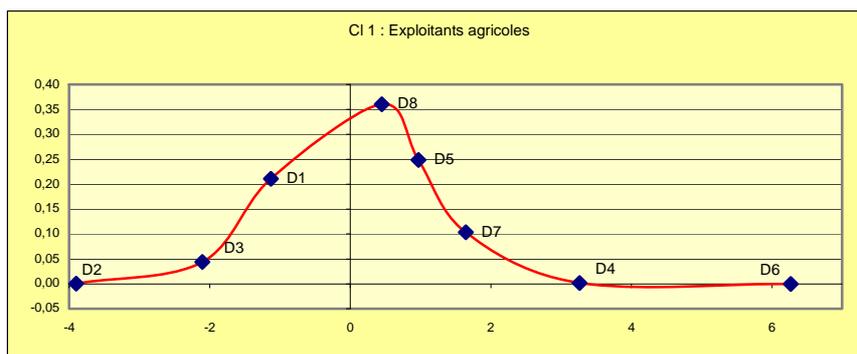
Rappelons encore une fois que la statistique ne se prononce pas sur l'essence des phénomènes, mais qu'elle les « découvre » et qu'elle les met en lumière ou en évidence pour mieux les analyser. La statistique « fait connaître » pour mieux « faire reconnaître ».

C'est ainsi qu'on peut mettre côte à côte tous les graphiques de même nature pour mieux comparer les phénomènes :

- 1) les graphiques en barre :



- 2) les graphiques en cloche :



Le graphique a une fonction géométrique : il permet de visualiser le phénomène pour mieux en percevoir les contours et les qualités inhérentes et de mieux orienter l'analyse, ou de l'orienter autant que possible de façon objective pour éviter de fausser les conclusions ou les interprétations.

Telle est la fonction essentielle de la TDR.

7. Conclusion provisoire

Plus que les résultats, c'est la méthode ici qui compte.

Cette méthode est d'autant plus pratique qu'il suffit d'activer la Macro pour analyser automatiquement un corpus de données.

Mais il convient de souligner que les conclusions seront d'autant plus justes et d'autant plus pertinentes que les relevés seront faits dans les règles de l'art.

La Macro, elle, traitera toujours avec efficacité les données et les tables de contingence, TDF et TDR, qui lui sont soumises, mais elle ne pourra en aucun cas rectifier quelque erreur que ce soit provenant d'un mauvais relevé.

Supposons encore qu'au lieu de traiter des classes sociales et des disciplines, l'on veuille analyser l'efficacité de certains traitements sur des malades, les résultats scolaires de plusieurs établissements, les données économiques des entreprises, des pays, la production laitière des régions, les pollutions des mers, la désertification des villes et des campagnes, etc., la statistique s'appliquerait toujours de la même façon, toujours avec la même rigueur et avec la même efficacité. Il dépend donc de tout un chacun d'en faire un bon usage, elle est vraiment à la portée de tous.

Dans le chapitre suivant nous allons aborder les Ajustements, les Corrélations et les ACP à partir des données de ce même corpus.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

2

Ajustement, Corrélation et ACP

par
André CAMLONG
Christine CAMLONG-VIOT

Ce deuxième chapitre est consacré à l'étude de l'Ajustement des variables, de la Corrélation et à l'Analyse en Composantes Principales (les ACP).

Nous utiliserons toujours comme support pédagogique le corpus présenté par G. Saporta à la page 151 des *Probabilités, Analyse des Données et Statistique*, tiré des *Données Sociales*, 3^e éd., INSEE, 1978.

Après un bref rappel des définitions et des formules mathématiques et statistiques utilisées, nous aborderons l'analyse des Ajustements et des Corrélations et nous exposerons la méthode de calcul de la Métrique R.

Comme dans le chapitre précédent, nous exploiterons les résultats fournis par la Macro qui est à la disposition de tout utilisateur intéressé.

Le lecteur qui aurait oublié les formules de calcul ou les théorèmes retenus par la statistique, ou qui aurait des difficultés à les comprendre, pourra utilement consulter des ouvrages spécialisés comme celui de G. Saporta déjà cité, *Probabilités, Analyse des Données et Statistique*. Paris : Édit. Technip, 1990, 493 p., ou celui de J. Grifone, *Algèbre linéaire*. Toulouse : Cépaduès-Éditions, 1990, 443 p.

Mais nous tenons d'ores et déjà à rassurer tous les lecteurs. C'est par la pratique qu'on comprendra le bien-fondé de ces calculs et qu'on se familiarisera avec les notions de transformée de Fourier, de lois de probabilité, d'inégalité de Cauchy-Schwarz, de relation de Pythagore, etc., ou avec les théorèmes de Craig et de Cochran, les lois de Laplace-Gauss, les méthodes de Box et Müller, la méthode polaire de Marsaglia, la *boxplot* de Tukey, etc., ou avec les notions de « *liaison stochastique* », de « *formes quadratiques* », etc.

La statistique est vraiment à la portée de tous.

1. Ajustement et Corrélation

L'*Ajustement* ou *Estimation* d'une variable est un mode de calcul qui consiste à estimer le degré de variation d'une variable en fonction d'une autre. C'est ce qu'on appelle encore la « *liaison stochastique* » : la connaissance de l'une servant de « cible » à la connaissance de

l'autre. Le degré de liaison qui est établi entre les deux variables s'appelle la *Corrélation* et s'exprime au moyen d'un coefficient dit coefficient de corrélation.

La variable connue X est dite *variable explicative* ou *prédicteur* et la variable inconnue Y est dite *variable expliquée* ou *critère*.

La droite d'estimation Y' de Y en X est la droite d'équation $Y' = b + aX$. C'est la droite dont la somme des carrés des distances aux divers points est minimale. Ces distances sont calculées parallèlement à OY , suivant la méthode des moindres carrés.

La définition de la droite d'estimation et le calcul de l'ajustement sont accompagnés d'un certain nombre de notions que nous allons définir et dont nous allons donner les formules appliquées au corpus retenu (emprunté à G. Saporta) :

Discipline / CI	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
D1 : Droit	80	6	168	470	236	145	166	16	305
D2 : Sciences Eco	36	2	74	191	99	52	64	6	115
D3 : Lettres	134	15	312	806	493	281	401	27	624
D4 : Sciences	99	6	137	400	264	133	193	11	247
D5 : Médecine-Dentaire	65	4	208	876	281	135	127	8	301
D6 : Pharmacie	28	1	53	164	56	30	23	2	47
D7 : Pluridisciplinaire	11	1	21	45	36	20	28	2	42
D8 : IUT	58	4	62	79	87	54	129	8	90

Les classes sociales qui sont portées dans les colonnes et numérotées de 1 à 9 suivant la nomenclature CI 1, CI 2, ..., CI 9, constituent les variables du corpus :

CI 1	Exploitants agricoles
CI 2	Salariés agricoles
CI 3	Patrons
CI 4	Professions libérales et Cadres supérieurs
CI 5	Cadres moyens
CI 6	Employés
CI 7	Ouvriers
CI 8	Personnel de service
CI 9	Autres

Rappel des formules et des définitions

- 1.1 les primitives X et Y . Ce sont les variables appariées. Comme l'un des couples ci-dessus, par exemple, CI 1 (X) et CI 2 (Y). Comment estimer Y en fonction du degré de connaissance de X ? Tel est le problème. Dans le tableau ci-dessus, chaque variable comprend 8 items (ou couples de disciplines), de telle sorte que : $X = X_1, X_2, \dots, X_8$ et $Y = Y_1, Y_2, \dots, Y_8$. Ou, en règle générale : $X = X_1, X_2, \dots, X_n$ et $Y = Y_1, Y_2, \dots, Y_n$. L'indice n désignant le nombre de couples (à savoir de lignes du tableau)

- 1.2 les moyennes marginales **Erreur ! Signet non défini.** \bar{X} et \bar{Y} sont les moyennes arithmétiques de la colonne
- 1.3 les valeurs centrées réduites x et y . Ce sont les différences entre les valeurs primitives et la moyenne pour chaque ligne : $x = (X_i - \bar{X})$ et $y = (Y_i - \bar{Y})$: il s'agit en fait de deux vecteurs que l'on considère dans l'espace R^n sous la forme d'un nuage de points
- 1.4 les coefficients d'estimation a (de Y en X) et a' (de X en Y). Ils sont donnés par les formules : $a = \frac{\sum xy}{\sum x^2}$ et $a' = \frac{\sum xy}{\sum y^2}$
- 1.5 le coefficient de corrélation r est le cosinus de l'angle formé par les deux vecteurs $x = (X_i - \bar{X})$ et $y = (Y_i - \bar{Y})$. Il est donné par la formule : $r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$
- 1.6 l'intervalle de variation de r est celui du cosinus : $-1 \leq r \leq +1$
- 1.7 le coefficient de détermination r^2 . Il est donné par la formule : $r^2 = a.a'$
- 1.8 l'équation de la droite d'estimation Y' : $Y' = \bar{Y} + a(X - \bar{X})$
- 1.9 les écarts-type σ_x et σ_y sont donnés par les formules : $\sigma_x = \sqrt{\frac{\sum x^2}{n}}$ et $\sigma_y = \sqrt{\frac{\sum y^2}{n}}$, n étant le nombre de couples ou de lignes.
- 1.10 *NB* : Nous retenons n comme ddl pour le traitement d'un corpus entier ou exhaustif et non la valeur $(n - 1)$ traditionnellement appliquée à un échantillon qui est par nature limité à moins de 30 éléments. Nous retrouverons cette valeur entière de n lors de l'étude de l'inertie.

On peut aisément appliquer ces formules au calcul des coefficients dans une feuille d'Excel, ou, plus simplement, faire appel aux fonctions prédéfinies d'Excel pour les exécuter.

2. Calculs pratiques

Appliquons ces formules au calcul des coefficients d'estimation et de corrélation effectués sur les deux premières variables du corpus de référence, X (Cl 1) et Y (Cl 2) :

	X	Y	x^2	y^2	xy
	80	6	260,0156	1,2656	18,1406
	36	2	777,0156	8,2656	80,1406
	134	15	4917,5156	102,5156	710,0156
	99	6	1233,7656	1,2656	39,5156
	65	4	1,2656	0,7656	-0,9844
	28	1	1287,0156	15,0156	139,0156
	11	1	2795,7656	15,0156	204,8906
	58	4	34,5156	0,7656	5,1406
somme	511	39	11306,8750	144,8750	1195,8750
moyenne	63,875	4,875			

			$a = 0,11$		
			$a' = 8,26$		
			$r = 0,934$		

D'où l'équation de la droite d'estimation Y' de Y en X :

$$Y' = a(X_i - \bar{X}) + \bar{Y}$$

ou

$$Y' = ax + \bar{Y}$$

D'où l'équation de la droite d'estimation Y' pour les deux premières variables du corpus :

$$Y' = 4,875 + 0,11 (X_i - 63,875)$$

Le problème de la droite d'estimation sera intégralement repris dans le chapitre suivant.

Le coefficient de corrélation $r = 0,934$ tend vers 1. Ce qui signifie que l'accroissement de Y est fortement lié à l'accroissement de X .

Comme l'intervalle de variation de r est identique à celui du *cosinus*, $-1 \leq r \leq +1$, on peut identifier la qualité des liaisons dans les différentes parties de l'intervalle :

- si $r = -1$, la liaison est *rigide négative*. Ce qui signifie que les accroissements de Y sont inversement proportionnels à ceux de X (X croît pendant que Y décroît ou inversement)
- si $-1 < r \leq 0$, la corrélation est négative et d'autant plus que r tend vers -1 . r est négatif en même temps que $\sum xy$; les coefficients a et a' , pentes de Y' et X' , sont également nuls
- si $r = 0$, les deux variables X et Y ne sont pas liées, ce qui ne signifie nullement qu'elles soient indépendantes (l'absence de liaison n'est pas synonyme d'indépendance)
- si $0 \leq r < 1$, la corrélation est positive et d'autant plus que r tend vers $+1$
- si $r = +1$, la liaison est *rigide positive*. Ce qui signifie que les accroissements de Y sont proportionnels aux accroissements de X

NB : un r modéré n'a qu'une très faible valeur explicative de la liaison.

Exemple : pour expliquer la variabilité de Y à hauteur de 50% de celle de X , il faut que le coefficient de corrélation $r = \pm 0,866$ (racine carrée de 0,75).

La *variance* de Y est ainsi expliquée à hauteur de r^2 % par celle de X (valeur du coefficient de détermination r^2) et la *partie résiduelle* ou *non expliquée* à hauteur de $(1 - r^2)$ %.

La variance de Y se décompose en $r^2\sigma_Y^2$, *partie expliquée* de la variation de Y et en $(1-r^2)\sigma_Y^2$, variance résiduelle ou *partie non expliquée* de Y , dans l'estimation de Y en X .

3. Matrice de Corrélation

La corrélation est intransitive, aussi les variables doivent-elles être appariées.

Pour calculer le coefficient de corrélation de 2 variables, on procède comme indiqué précédemment, ou bien on utilise la fonction COEFFICIENT.CORRELATION, prédéfinie d'Excel, et on confectionne la matrice triangulaire :

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
CI 1	1								
CI 2	0,934	1							
CI 3	0,884	0,893	1						
CI 4	0,721	0,679	0,933	1					
CI 5	0,935	0,930	0,983	0,886	1				
CI 6	0,943	0,961	0,976	0,843	0,991	1			
CI 7	0,959	0,995	0,893	0,689	0,938	0,960	1		
CI 8	0,927	0,977	0,892	0,677	0,914	0,957	0,969	1	
CI 9	0,910	0,952	0,982	0,857	0,987	0,996	0,946	0,949	1

En fixant le seuil limite de signification à $r = 0,866$, le taux de corrélation des variables appariées est immédiatement saisi.

Toute valeur inférieure à 0,866 indique une mauvaise liaison entre les variables appariées. C'est le cas de la liaison de CI 4 avec CI 2 ($r = 0,721$), avec CI 3 ($r = 0,679$), avec CI 7 ($r = 0,689$) et avec CI 8 ($r = 0,677$) qui est médiocre ou mauvaise.

4. Matrice de Détermination

La matrice des coefficients de détermination r^2 , complémentaire de la matrice de corrélation, propose une lecture directe des taux de liaison :

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
CI 1	1								
CI 2	0,873	1							
CI 3	0,781	0,797	1						
CI 4	0,519	0,462	0,871	1					
CI 5	0,875	0,865	0,966	0,786	1				
CI 6	0,888	0,923	0,953	0,711	0,983	1			
CI 7	0,920	0,989	0,797	0,475	0,880	0,923	1		
CI 8	0,859	0,954	0,796	0,459	0,835	0,916	0,939	1	
CI 9	0,829	0,906	0,965	0,735	0,974	0,992	0,895	0,901	1

La liaison entre CI 1 (X) et CI 2 (Y) est expliquée à hauteur de 87,3%.
entre CI 1 (X) et CI 3 (Y) à hauteur de 78,1%.
entre CI 1 (X) et CI 4 (Y) à hauteur de 51,9%.

La variance de CI 6, par exemple, est expliquée à hauteur de 88,8% par celle de CI 1,
à hauteur de 92,3% par celle de CL 2,
à hauteur de 95,3% par celle de CI 3,
à hauteur de 98,3 par celle de CI 5,
à hauteur de 99,2% par celle de CI 9.

A l'inverse, la variance de CI 4 n'est expliquée qu'à hauteur de 51,9% par celle de CI 1,
à hauteur de 46,2% par celle de CI 2,
à hauteur de 47,5% par celle de CI 7,
à hauteur de 45,9% par celle de CI 8. Ce qui est très faible.

La statistique remplit parfaitement son rôle, qui n'est pas de se prononcer sur l'essence des phénomènes, mais de les déceler et de les identifier pour en faciliter l'analyse et l'interprétation.

La statistique est un outil de description et d'aide à l'interprétation.

5. Les valeurs marginales de Corrélacion et la métrique R

Les valeurs marginales de la matrice de corrélation forment un vecteur analytique, dans un même espace fermé, où les variables se positionnent les unes par rapport aux autres de façon à la fois indépendante et hiérarchique, permettant de dégager les profils des liaisons sous forme de nuage de points.

5.1 Le vecteur des valeurs marginales de corrélation

La valeur marginale de corrélation est le vecteur de la moyenne des corrélations pour chaque variable de la matrice carrée. Cet \bar{r} , nous l'appellerons tout simplement r , valeur primordiale de la *métrique R*.

Il faut dans un premier temps transformer la matrice triangulaire en matrice carrée, une matrice symétrique et positive :

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
CI 1	1	0,934	0,884	0,721	0,935	0,943	0,959	0,927	0,910
CI 2	0,934	1	0,893	0,679	0,930	0,961	0,995	0,977	0,952
CI 3	0,884	0,893	1	0,933	0,983	0,976	0,893	0,892	0,982
CI 4	0,721	0,679	0,933	1	0,886	0,843	0,689	0,677	0,857
CI 5	0,935	0,930	0,983	0,886	1	0,991	0,938	0,914	0,987
CI 6	0,943	0,961	0,976	0,843	0,991	1	0,960	0,957	0,996
CI 7	0,959	0,995	0,893	0,689	0,938	0,960	1	0,969	0,946
CI 8	0,927	0,977	0,892	0,677	0,914	0,957	0,969	1	0,949
CI 9	0,910	0,952	0,982	0,857	0,987	0,996	0,946	0,949	1

et, dans un deuxième temps, calculer la valeur moyenne du coefficient de corrélation de chaque variable (colonne) :

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
<i>r</i>	0,913	0,924	0,937	0,810	0,952	0,959	0,928	0,918	0,953

Ce vecteur n'est dépourvu ni de signification ni d'intérêt :

- 1) les valeurs de *r* sont des valeurs réelles et non des valeurs hypothétiques
- 2) chaque *r* est un *cosinus* qui positionne chaque variable comme un point du nuage dans un espace polaire
- 3) chaque point du nuage se définit par une abscisse, *le cosinus r* (cosinus connu), et une ordonnée, *le sinus ρ* (calculé par déduction)
- 4) les valeurs de *r* sont les valeurs primordiales utilisées par la *métrique R* pour calculer les profils hiérarchiques projetés dans un espace euclidien

5.2 La métrique R

Les mesures de la métrique R ne sont pas des mesures de grandeur mais des mesures de positionnement hiérarchique des variables. Grâce à ces « mesures relatives », on peut observer les profils des distributions sous le meilleur angle : ce sont les ACP (Analyses en Composantes Principales).

La métrique R (ainsi appelée par référence au coefficient de corrélation *r*) va proposer 4 séries de couples de coordonnées permettant de situer chaque point du nuage (qui est la réduction de chaque variable à sa plus simple expression) dans un espace polaire fermé et d'y opérer les meilleures projections et rotations et en dégager (par endomorphisme) les images les plus représentatives de la corrélation que les variables peuvent entretenir entre elles.

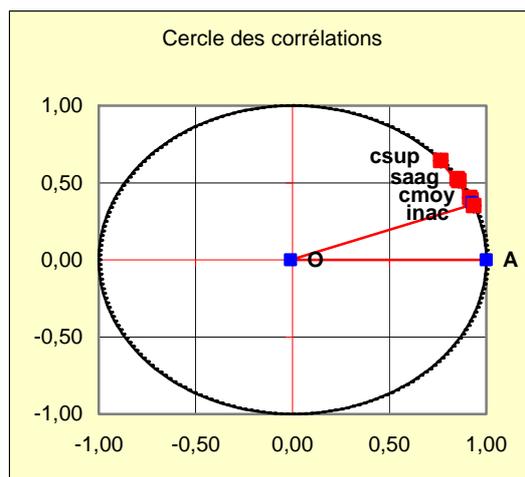
5.2.1 Première paire de coordonnées : *cosinus r* et *sinus ρ*

Suivant la relation de Pythagore, $\sinus \rho = \sqrt{1 - r^2}$.

D'où les coordonnées primordiales de chaque variable, laquelle est réduite à sa plus simple expression, celle d'un point du nuage.

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
<i>r</i>	0,913	0,924	0,937	0,810	0,952	0,959	0,928	0,918	0,953
<i>ρ</i>	0,409	0,381	0,349	0,587	0,307	0,285	0,373	0,396	0,302

Le nuage de points se projette bien dans un cercle de centre O et de rayon 1, de telle sorte que chaque point, en fonction de l'inégalité de Cauchy-Schwarz, se projette à l'intérieur du cercle, dans l'un des quadrants (ici le quadrant supérieur droit positif).

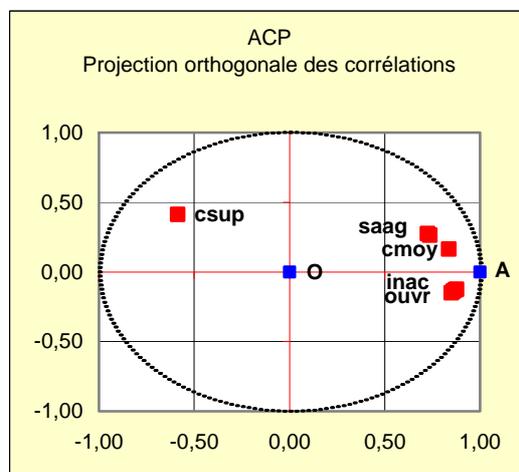


Le cercle est un cercle de centre O et de rayon 1. Sur la feuille d'Excel, le curseur donne immédiatement les coordonnées de chaque point.

L'image montre que l'ensemble des points est centré, bien que l'un d'eux se détache du groupe (celui de la variable Cl 4).

L'image reflète parfaitement l'homogénéité du corpus analysé.

La projection orthogonale (par le carré du *cosinus* et du *sinus*) permet de situer la matrice dans le cercle :



Cette projection est fondamentale pour l'étude de l'ACP.

On peut également calculer l'angle que la droite qui part de l'origine O, centre du cercle, et passe par le point du nuage. Pour cela, utiliser la fonction ACOS, qui donne la valeur de l'arc, que l'on transforme ensuite en degrés, par exemple, grâce à la fonction DEGRES.

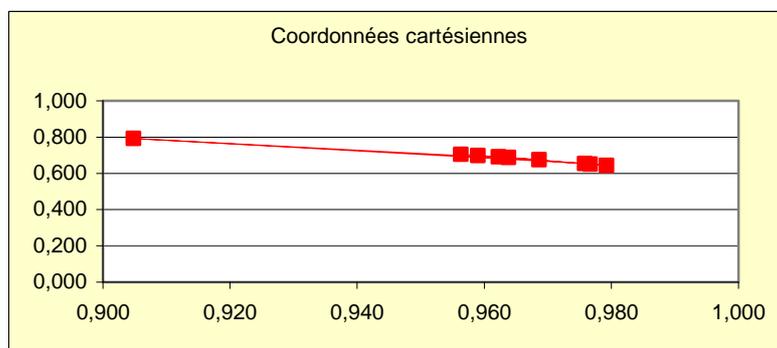
Prenons l'exemple de Cl4 : $\cosinus\ 0,810$; $arc = 0,627$; $angle = 35^{\circ}54'$. C'est l'angle le plus ouvert.

L'angle le plus fermé, celui de Cl4 : $\cosinus = 0,956$; $arc = 0,287$; $angle = 16^{\circ}28'$.

5.2.2 Deuxième paire de coordonnées cartésiennes x et y

Les coordonnées cartésiennes $x = \frac{r+1}{2}$ et $y = \frac{\rho+1}{2}$ permettent de situer chaque point dans un espace euclidien élargi ou ouvert.

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
x	0,956	0,962	0,969	0,905	0,976	0,979	0,964	0,959	0,977
y	0,704	0,691	0,674	0,793	0,654	0,642	0,687	0,698	0,651



L'image fait ressortir l'étalement des points du nuage.

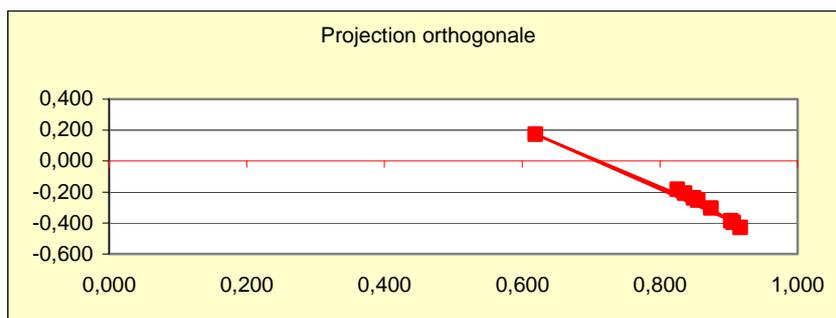
On remarque ici le regroupement de 8 éléments et l'égaré du 9^{ème} (CI 4).

5.2.3 Troisième paire de coordonnées r' et ρ'

Les coordonnées $r' = 2r - 1$ et $\rho' = 2\rho - 1$ permettent de projeter l'image des points dans un espace orthogonal, suivant une transformation appelée isométrie vectorielle gauche.

L'angle de rotation est donné par la formule $TrA = 2 \cos a - 1 (= 2r - 1)$, le déterminant $\det A = -1$.

	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
r'	0,825	0,849	0,875	0,619	0,903	0,917	0,855	0,836	0,907
ρ'	-0,182	-0,237	-0,303	0,174	-0,386	-0,430	-0,253	-0,207	-0,396



La projection orthogonale prend en compte les niveaux de la distribution.

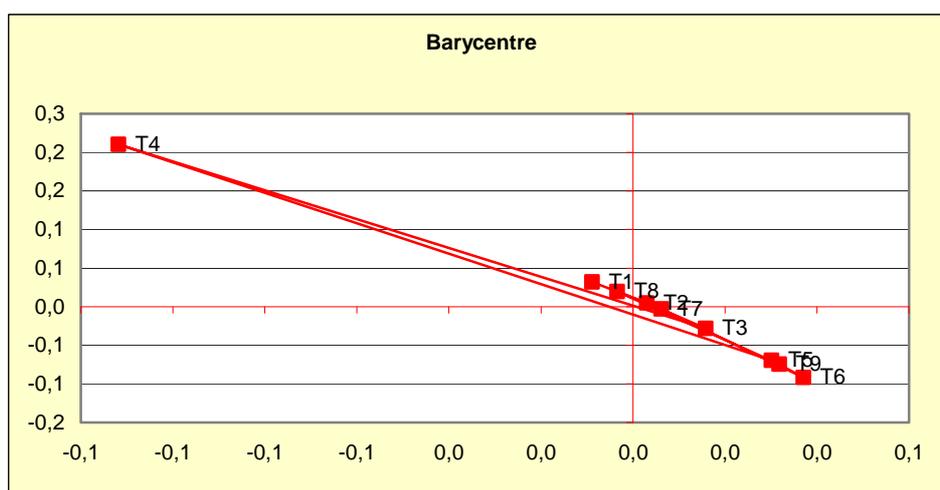
On voit ici 8 éléments regroupés dans la partie négative (à sinus négatif) et 1 élément (CI 4) qui est isolé dans la partie positive (à sinus positif).

5.2.4 Quatrième paire de coordonnées r'' et ρ''

Les coordonnées $r'' = r - \bar{r}$ et $\rho'' = \rho - \bar{\rho}$ permettent de faire une projection oblique avec rotation autour du centre de gravité (barycentre).

Il est aisé de calculer r'' et ρ'' sachant que \bar{r} et $\bar{\rho}$ sont les moyennes vectorielles de r et ρ .

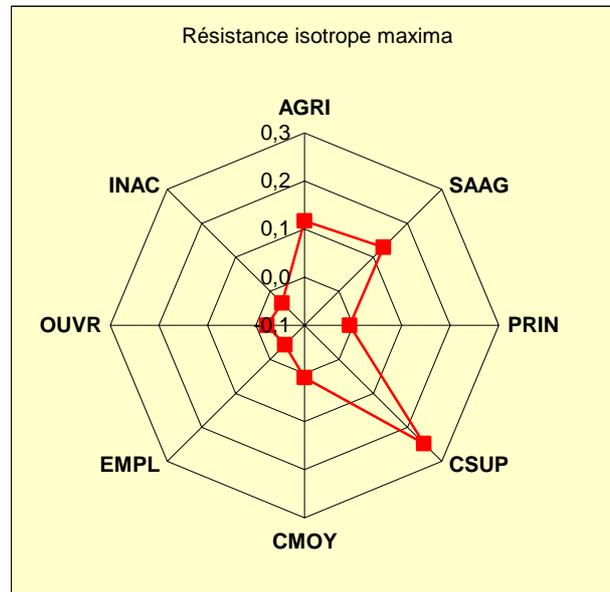
	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
r''	-0,009	0,003	0,016	-0,112	0,030	0,037	0,006	-0,003	0,032
ρ''	0,032	0,005	-0,028	0,210	-0,069	-0,092	-0,003	0,020	-0,075



Cette image est la meilleure image possible de la distribution : c'est l'image d'inertie de la distribution autour du centre de gravité.

On peut avoir cette sous forme de radar qui sera une image de meilleure facture lorsqu'on veut mettre en évidence la résistance des éléments en fonction du *sinus* qui « qualifie » la

variance résiduelle $(1-r^2)\sigma_y^2$ ou la partie non expliquée de la liaison, sachant que la partie expliquée $r^2\sigma_y^2$ est « qualifiée » par le *cosinus*.



L'image du radar focalisée sur le sinus ρ'' met ici en évidence l'élément excentré (de CI 4), l'élément qui résiste à la liaison (le sinus ρ'' étant le coefficient de résistance d'inertie maxima).

La hiérarchie des variables peut encore être mise en relief au moyen de structures arborescentes ou de dendrogrammes. La Macro le fait automatiquement.

6. La Macro, les ACP et les graphes

Pourquoi se compliquer la vie quand on peut se la simplifier ? La solution, c'est la macro qui exploite la base de données et fait automatiquement calculs et graphes.

Voici la matrice de calcul de la métrique R relative au corpus retenu :

	<i>moyenne</i>	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
r	0,921	0,913	0,924	0,937	0,810	0,952	0,959	0,928	0,918	0,953
ρ	0,377	0,409	0,381	0,349	0,587	0,307	0,285	0,373	0,396	0,302
x	0,961	0,956	0,962	0,969	0,905	0,976	0,979	0,964	0,959	0,977
y	0,688	0,704	0,691	0,674	0,793	0,654	0,642	0,687	0,698	0,651
		0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
r'	0,843	0,825	0,849	0,875	0,619	0,903	0,917	0,855	0,836	0,907
ρ'	-0,247	-0,182	-0,237	-0,303	0,174	-0,386	-0,430	-0,253	-0,207	-0,396

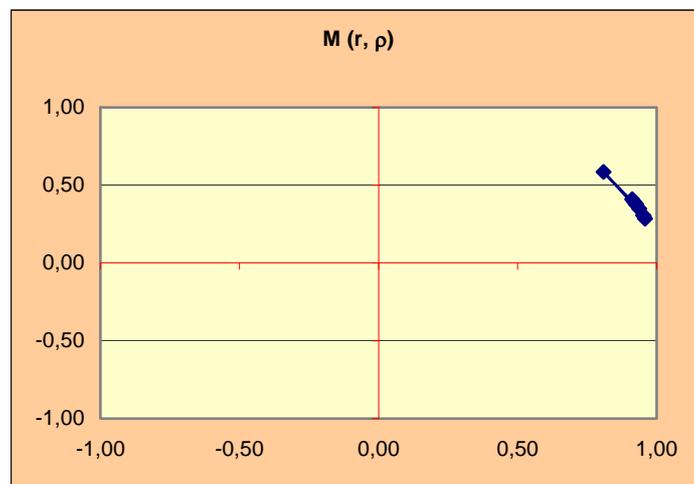
r''	0,000	-0,009	0,003	0,016	-0,112	0,030	0,037	0,006	-0,003	0,032
ρ''	0,000	0,032	0,005	-0,028	0,210	-0,069	-0,092	-0,003	0,020	-0,075

Cette grille reproduit les résultats des calculs effectués automatiquement.

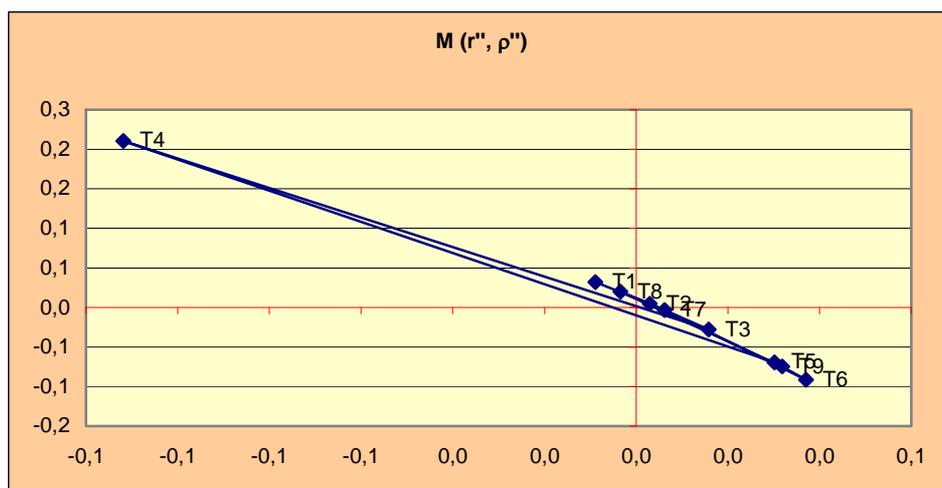
Dans la colonne dite « moyenne » figurent les coordonnées des centres de gravité du nuage de points formé par les différentes paires de coordonnées.

La Macro donne automatiquement les 4 types de graphes qui suivent :

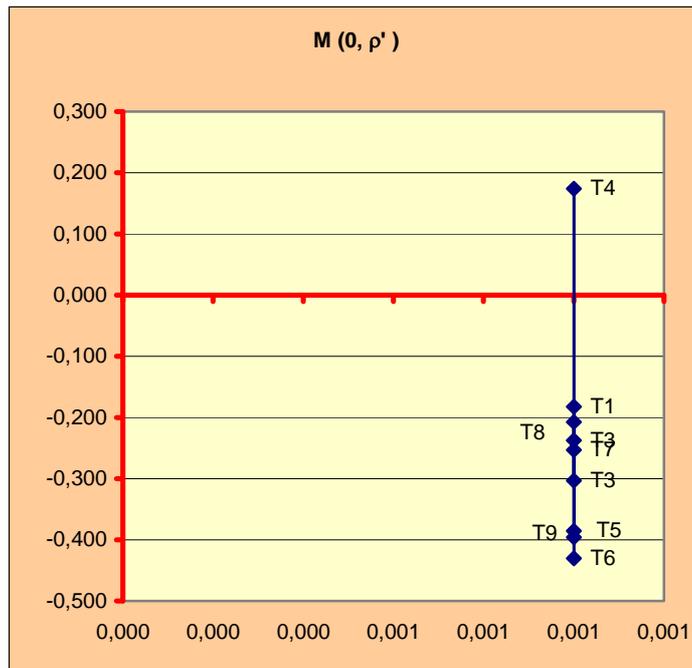
1) le graphe du « cercle carré » relatif aux données primordiales du corpus :



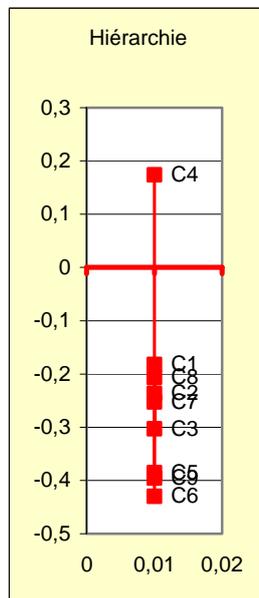
2) le graphe représentant le barycentre du nuage :



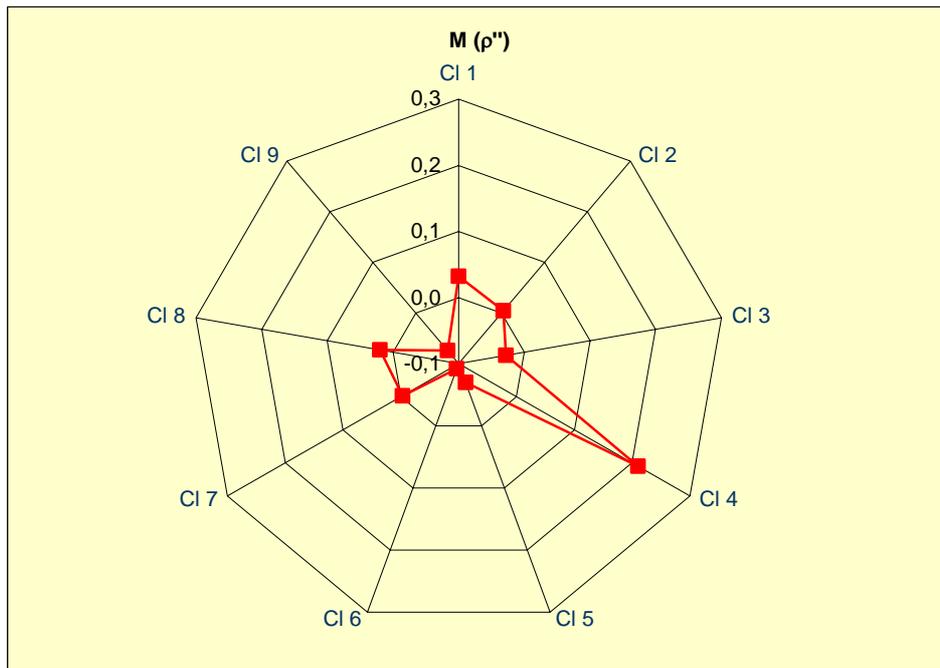
3) le dendrogramme vertical (ou horizontal) que l'on fait instantanément :



Excel offre toutes les possibilités de modifier et d'adapter les graphiques aux besoins de la présentation. Ces graphiques qui appartiennent tous à l'univers géométrique, n'ont d'autre fonction que de « faire voir » et « faire connaître » pour mieux « faire reconnaître ».



4) le radar mettant en exergue les éléments qui « résistent » à la liaison (*sinus*) :



Les calculs et les graphes ne sont reproduits ici qu'à titre d'exemple, ils ne sont donc pas accompagnés de commentaires.

La Macro est accompagnée d'un ouvrage qui en décrit l'utilisation et les potentialités.

Mais cette Macro comprend également une autre page, la huitième, où se font les calculs de régression et les graphes appropriés. Cette description va faire l'objet du troisième chapitre. La Régression est l'une des phases les plus riches et les plus utiles de l'analyse statistique, puisqu'elle débouche directement sur l'Analyse Factorielle Discriminante (AFD).

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

3

La Régression linéaire

par
André CAMLONG
Christine CAMLONG-VIOT

Ce troisième chapitre est consacré à l'étude de la Régression linéaire laquelle se définit en tant qu'Analyse Linéaire Discriminante, mais qui est, *de facto*, une Analyse Factorielle Discriminante (AFD).

Régression, Estimation, Ajustement, Prévision, Prédiction, Prédicteur, Critère, ce sont tous, on le sait, des termes descriptifs de la liaison stochastique entre 2 variables appariées X et Y : Y est le *critère* estimé en fonction du *prédicteur* X suivant l'équation $Y' = b + aX$ de la droite de régression Y' : Y' est la droite d'estimation de Y en X .

Trois cas de régression linéaire :

1. la régression linéaire simple est la liaison stochastique entre 2 variables, le *critère* $Y = (Y_1, Y_2, Y_3, \dots, Y_n)$ et le *prédicteur* $X = (X_1, X_2, X_3, \dots, X_n)$
2. la régression linéaire multiple est la liaison stochastique entre le *critère* $Y = (Y_1, Y_2, Y_3, \dots, Y_n)$, l'une des variables du corpus, en fonction du *prédicteur marginal* X , la variable $p + 1$, somme des p variables du corpus.
3. la régression linéaire factorielle ou vectorielle est affaire de discrimination ou de lemmatisation, c'est l'Analyse Factorielle Discriminante (AFD) par excellence.

Comme dans les chapitres précédents, nous utiliserons le corpus présenté par G. Saporta à la page 151 des *Probabilités, Analyse des Données et Statistique*, Paris : Édit. Technip, 1990, tiré des *Données Sociales*, 3^e éd., INSEE, 1978.

Après un bref rappel des définitions et des formules, nous suivrons pas à pas le déroulement des opérations effectuées par la Macro pour en comprendre les modes opératoires, en saisir la finalité, en déterminer les moments d'inertie et en visualiser les résultats sur des graphiques appropriés.

1. Régression, Ajustement, Estimation, Prédiction, Liaison stochastique

1.1 Définitions

La droite de régression de Y en X est la droite d'estimation Y' d'équation $Y' = b + aX$:

$$Y' = a(X_i - \bar{X}) + \bar{Y}$$

ou

$$Y' = ax + \bar{Y}$$

- X est dite variable explicative ou *prédicteur* : elle est *déterminante*
- Y est dite variable expliquée ou *critère* : elle est *déterminée* ou *estimative*
- Y' est dite droite d'estimation de Y en X produite par l'ajustement des 2 variables en fonction d'une *liaison stochastique*.

La droite d'estimation Y' de Y en X (dite de régression dans les travaux de Galton) est la droite définie par la somme des carrés des distances minimales $e = Y - Y'$. Ces distances, appelées *résidus*, sont calculées parallèlement à OY , suivant la méthode des *moindres carrés*. (Voir Morice et Chartier, *Méthode statistique*, p. 275).

En fonction du *prédicteur* X , on espère définir le *critère* Y sinon d'une façon parfaite, du moins d'une façon proche de la réalité. C'est le sens qui est donné à *liaison stochastique*, calcul qui prend X pour « *cible* » (suivant l'étymologie grecque de *stochos*). C'est en fonction du degré de connaissance de X que l'on va, *non pas déterminer exactement* Y , *mais en fournir une estimation*.

1.2 Matrice des données

Plutôt que d'enfoncer des portes ouvertes, nous allons travailler concrètement sur la base des données de G. Saporta que nous avons utilisée dans les chapitres précédents :

Discipline / CI	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
D1 : Droit	80	6	168	470	236	145	166	16	305
D2 : Sciences Eco	36	2	74	191	99	52	64	6	115
D3 : Lettres	134	15	312	806	493	281	401	27	624
D4 : Sciences	99	6	137	400	264	133	193	11	247
D5 : Médecine-Dentaire	65	4	208	876	281	135	127	8	301
D6 : Pharmacie	28	1	53	164	56	30	23	2	47
D7 : Pluridisciplinaire	11	1	21	45	36	20	28	2	42
D8 : IUT	58	4	62	79	87	54	129	8	90

Les classes sociales qui sont portées dans les colonnes et numérotées de 1 à 9 suivant la nomenclature Cl 1, Cl 2, ..., Cl 9, constituent les variables du corpus :

Cl 1	Exploitants agricoles
Cl 2	Salariés agricoles
Cl 3	Patrons
Cl 4	Professions libérales et Cadres supérieurs
Cl 5	Cadres moyens
Cl 6	Employés
Cl 7	Ouvriers
Cl 8	Personnel de service
Cl 9	Autres

Nous retiendrons les deux premières variables X (Cl 1) et Y (Cl 2) : $X = X_1, X_2, \dots, X_8$ et $Y = Y_1, Y_2, \dots, Y_8$ (chaque variable comprend 8 items ou couples de disciplines).

En règle générale : $X = X_1, X_2, \dots, X_n$ et $Y = Y_1, Y_2, \dots, Y_n$. Et l'indice n désigne le nombre de couples ou de lignes du tableau.

Rappel. L'indice n , c'est aussi le nombre de degrés de liberté (ddl) d'une population complète ; c'est aussi la trace de la matrice de corrélation symétrique et positive.

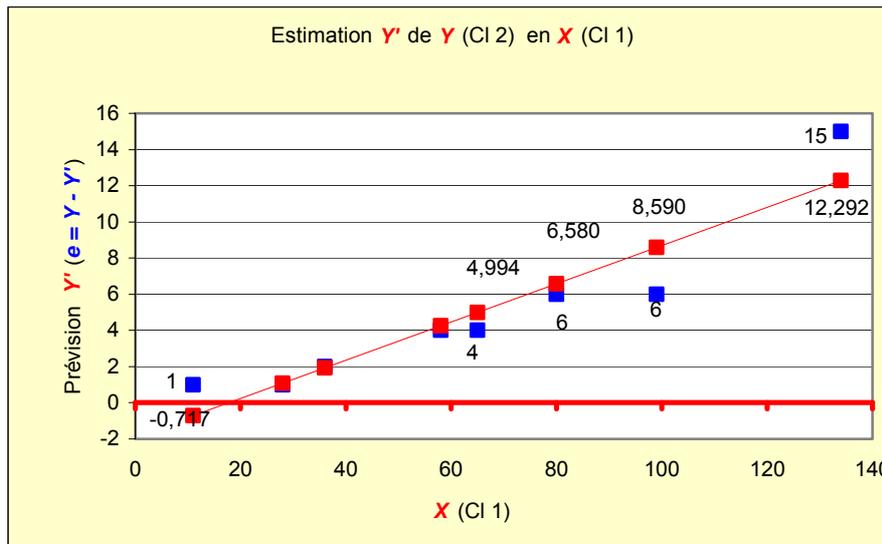
S'agissant d'un échantillon de « petite taille » tiré au sort (selon les normes de Tippett, par exemple), le ddl est réduit à $(n - 1)$.

En conséquence, nous considérerons que le corpus de 10.000 individus retenu par l'INSEE et par Saporta tient plus de la population exhaustive que de l'échantillon à proprement parler.

Ajoutons, pour être complet, que l'on prenne pour dividende n ou $(n-1)$, lorsque la population est élevée, le résultat est sensiblement le même.

1.3 Observations

Observons le graphique de l'estimation Y' de Y (Cl 2) en X (Cl 1), les deux premières variables du corpus précité :



Les *résidus* $e = Y - Y'$ qui préfigurent la *distance des moindres carrés* constituent le nuage de points de Y (en bleu) autour de l'estimation Y' (points alignés en rouge).

En fonction des ordonnées des points du nuage, la distance $e = Y - Y'$ est excédentaire ou déficitaire (suivant le point bleu au-dessus ou au-dessous du point rouge).

La *liaison stochastique* se réduit à une bande d'autant plus étroite autour de Y' que le coefficient de corrélation r est voisin de ± 1 . Si la liaison est *rigide*, le nuage de points tend à se confondre avec la droite d'estimation.

Le calcul de la droite d'estimation Y' doit toujours s'accompagner du calcul du coefficient de corrélation r qui en précise la signification.

Dans le graphique ci-dessus on voit que 3 points sur 8 se détachent nettement de la droite d'estimation.

Le coefficient de corrélation entre les deux variables appariées est proche de 1 ($r = 0,934$). Cela signifie qu'il y a une forte *liaison stochastique* entre X et Y . (Voir chap. 2, 2)

Quelle en est la signification ? Quelle en est la valeur descriptive ? Quelle en est la valeur explicative ? C'est ce que nous allons découvrir et essayer de comprendre.

La statistique ne se prononce pas, ou rarement, sur l'essence des phénomènes observés, elle les « *fait connaître* » pour mieux les « *faire reconnaître* ».

La statistique est un instrument de description et d'aide à l'interprétation.

1.4 Rappel des formules

L'équation de la droite d'estimation Y' est donnée par la formule :

$Y' = a(X_i - \bar{X}) + \bar{Y}$ <p>ou</p> $Y' = ax + \bar{Y}$

- 1) a est le coefficient d'estimation ou de régression de Y en X : $a = \frac{\sum xy}{\sum x^2}$
- 2) r est le coefficient de corrélation des 2 variables : $r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$ ou cosinus r de l'angle θ formé par les 2 vecteurs $x = (X_i - \bar{X})$ et $y = (Y_i - \bar{Y})$.
- 3) l'intervalle de variation de r est celui du cosinus : $-1 \leq r \leq +1$
- 4) σ_x et σ_y sont les écarts-type des 2 vecteurs : $\sigma_x = \sqrt{\frac{\sum x^2}{n}}$ et $\sigma_y = \sqrt{\frac{\sum y^2}{n}}$, n étant le nombre de couples ou de lignes.

À partir de ces formules il est aisé d'exécuter les calculs directement sur une feuille d'Excel ou de les simplifier en utilisant les fonctions prédéfinies.

2. Calculs pratiques

Rappel des calculs effectués dans l'article précédent pour définir l'équation de la droite d'estimation Y' , en fonction des 2 premières variables du corpus de référence, X (CI 1) et Y (CI 2) :

	X	Y	x^2	y^2	xy
	80	6	260,0156	1,2656	18,1406
	36	2	777,0156	8,2656	80,1406
	134	15	4917,5156	102,5156	710,0156
	99	6	1233,7656	1,2656	39,5156
	65	4	1,2656	0,7656	-0,9844
	28	1	1287,0156	15,0156	139,0156
	11	1	2795,7656	15,0156	204,8906
	58	4	34,5156	0,7656	5,1406
somme	511	39	11306,8750	144,8750	1195,8750
moyenne	63,875	4,875			
		$a = 0,11$			
		$a' = 8,26$			
		$r = 0,934$			

D'où l'équation de la droite d'estimation Y' de Y en X :

$$Y' = a(X_i - \bar{X}) + \bar{Y} \text{ ou } Y' = ax + \bar{Y}$$

D'où Y' de Y en X pour le couple choisi :

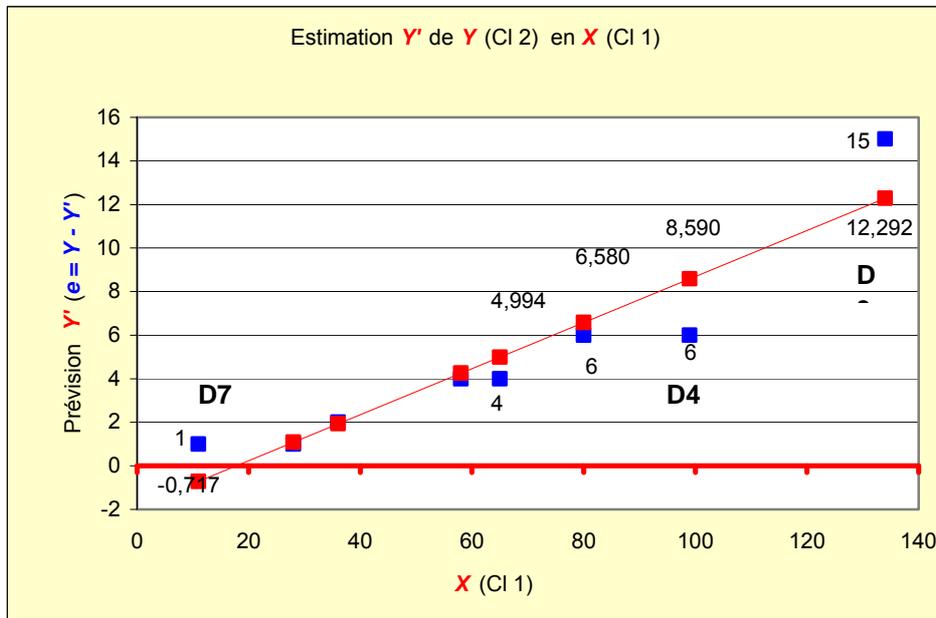
$$Y' = 4,875 + 0,11 (X_i - 63,875)$$

En répétant ce calcul pour les n couples (8 lignes) des 2 variables appariées, la droite d'estimation Y' est immédiatement connue. Les liaisons stochastiques sont identifiées et les résidus déchiffrés.

Les calculs de régression sont facilités dès lors que l'on utilise « TENDANCE », la fonction prédéfinie d'Excel étendue à toute la plage :

X	Y	Y'	Discipline
80	6	6,580	D1 : Droit
36	2	1,927	D2 : Sciences Eco
134	15	12,292	D3 : Lettres
99	6	8,590	D4 : Sciences
65	4	4,994	D5 : Médecine-Dentaire
28	1	1,081	D6 : Pharmacie
11	1	-0,717	D7 : Pluridisciplinaire
58	4	4,254	D8 : IUT

D'où la droite d'estimation Y' (en rouge) et le nuage de points de Y en X :



Confection du graphique du nuage de points en 2 temps :

- 1) tracer le nuage de points de X et de Y
- 2) faire glisser Y' dans le graphique

Une bonne lecture du graphique vaut les meilleures explications du monde, à condition de tenir compte l'échelle qui est en marge.

On voit que 3 des 8 points du nuage se singularisent par un écart e ($e = Y - Y'$) positif ou négatif entre Y et Y' , aux 2 extrémités de la droite : **D3** et **D7** accusent une surcharge excédentaire, et **D4** accuse au contraire un déficit.

Quelle en est la signification ? Dès l'instant que les facteurs incriminés sont identifiés, la réponse est immédiate. On peut se prononcer sur les phénomènes décrits, mais on ne peut pas se prononcer sur les motivations. L'essence du problème ne relève pas de la statistique. On voit des préférences nettement marquées pour certaines disciplines.

Pour couper court à tout commentaire (qui n'est pas de mise ici), nous allons poursuivre la décomposition de la variance et la transformation opérée par les filtres linéaires qui produisent les spectres de densité et engendrent les moments d'inertie

3. Profils, filtres et spectres de densité

La variance totale, en fonction du principe d'additivité, se décompose en variance résiduelle et en variance contrôlée :

$$\boxed{\text{Variance totale} = \text{Variance résiduelle} + \text{Variance contrôlée}}$$

D'où la formule simplifiée :

$$\boxed{\sum (Y - \bar{Y})^2 = \sum (Y - Y')^2 + \sum (Y' - \bar{Y})^2}$$

D'où la formule déjà décrite dans le chapitre précédent :

$$\boxed{\sigma_Y^2 = \rho^2 \sigma_{Y'}^2 + r^2 \sigma_Y^2}$$

D'où la racine de décomposition utilisée dans les calculs de l'estimation Y' :

$$\boxed{(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})}$$

Cette formule permet de contrôler l'exactitude des calculs de chaque vecteur ou de chaque ligne d'analyse.

Elle permet en outre de vérifier que les variables de décomposition sont bien des variables indépendantes dans la mesure où la somme des valeurs de chaque colonne est nulle (conformément aux théorèmes de Craig et de Cochran).

Ces variables indépendantes obéissent aux lois centrales de Laplace-Gauss, LG (0 ; 1) et suivent les lois du χ^2 .

Les filtres quadratiques transforment logiquement la variance décomposée en vecteurs linéaires, dont la puissance des spectres souligne les qualités inhérentes aux vecteurs gaussiens.

Les filtres sont fournis par les formules canoniques de calcul des valeurs centrées réduites.

Les résultats sont immédiatement lisibles, puisqu'il s'agit d'écartés centrés et réduits qui expriment la dispersion du nuage de points autour de l'axe des abscisses suivant les limites classiques de probabilité :

1. V_t , la valeur centrée réduite totale : $V_t = \frac{Y - \bar{Y}}{\sigma_Y}$ avec σ_Y (écart-type de Y)

2. V_r , la valeur centrée réduite résiduelle : $V_r = \frac{Y - Y'}{u}$ avec $u = \rho \cdot \sigma_Y$ (résidu quadratique moyen)

3. V_c , la valeur centrée réduite contrôlée : $V_c = \frac{Y' - \bar{Y}}{c}$ avec $c = r \cdot \sigma_Y$ (contrôle quadratique moyen)

D'où le vecteur linéaire de transformation spectrale :

$$\boxed{V_t = V_r + V_c}$$

Or, les vecteurs gaussiens de décomposition et de transformation de la variance sont des vecteurs indépendants, qui suivent les lois de Laplace-Gauss, LG (0 ;1), selon les principes d'additivité du χ^2 , conformément aux théorèmes de Craig et de Cochran, qui ne sont rien moins qu'une variante du théorème de Pythagore, appliqués à la décomposition d'un vecteur (une forme quadratique du système linéaire) :

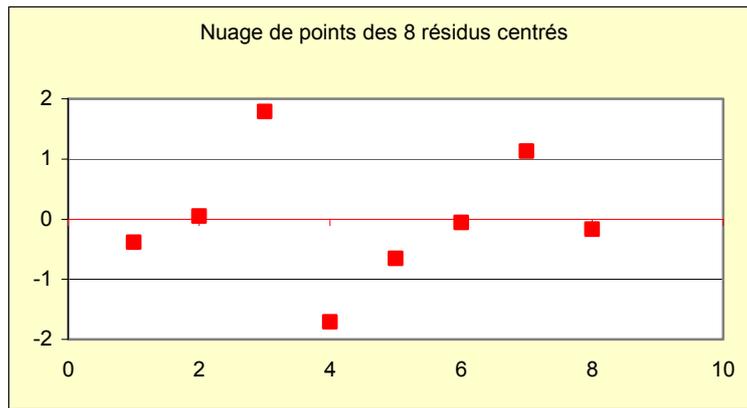
$$1. \sum V_t = 0$$

$$2. \sum V_r = 0$$

$$3. \sum V_c = 0$$

La transformation du vecteur gaussien en spectre de densité repose sur les propriétés d'additivité du χ^2 et garantit l'*homoscédasticité* des résidus e ($e = Y - Y'$), qui focalisent toute la puissance de l'analyse spectrale.

Dans l'*homoscédasticité*, le nuage de points ne laisse apparaître aucune tendance :



La distance quadratique dt du vecteur au centre de gravité de l'espace de projection orthogonale suit les lois du χ^2 de Fisher :

$$dt = \sqrt{\frac{1}{2} [(Vc^2 + Vr^2) - Vt^2]}$$

Calculs automatiquement effectués par la Macro.

NB : Le spectre suit une loi de χ^2 , un χ^2 de R (Y'), dont le nombre de degrés de liberté est le nombre de vecteurs (ou de lignes) qui composent la matrice d'estimation :

$$\langle f(x), f(y) \rangle = \frac{1}{2} (\|x + y\|^2 - \|x\|^2 - \|y\|^2) = \langle x, y \rangle$$

D'où la notion de distance quadratique du χ^2 de R (Y').

La distance est une mesure stable et sans dimension qui rend compte de la hiérarchie des individus.

C'est aussi un moment d'inertie du spectre.

4. Les calculs effectués par la Macro

Pour calculer Y' , reprenons les 2 premières variables du corpus, X (C1 1) et Y (C12) et suivons pas à pas les calculs effectués à la page 8 de la Macro.

1) Les primitives et l'estimation

Les 3 vecteurs fondamentaux sont les primitives X et Y et l'estimation Y' . Les primitives X et Y sont importées, l'estimation Y' est calculée :

<i>X</i>	<i>Y</i>	<i>Y'</i>	<i>Discipline</i>
80	6	6,580	D1 : Droit
36	2	1,927	D2 : Sciences Eco
134	15	12,292	D3 : Lettres
99	6	8,590	D4 : Sciences
65	4	4,994	D5 : Médecine-Dentaire
28	1	1,081	D6 : Pharmacie
11	1	-0,717	D7 : Pluridisciplinaire
58	4	4,254	D8 : IUT
511	39	39	Total

2) Les paramètres et les filtres

La Macro calcule automatiquement les paramètres, les coefficients et les valeurs quadratiques moyennes à partir du cosinus r : la moyenne \bar{Y} et l'écart-type σ_Y , les coefficients de corrélation r et de détermination r^2 , le sinus ρ et les écarts quadratiques moyens u et c (écarts-type de Y liés par X).

\bar{Y}	r	r^2	ρ	σ_Y	$u = \rho \cdot \sigma_Y$	$c = r \cdot \sigma_Y$
4,875	0,934	0,873	0,356	4,256	1,516	3,976

On vérifiera l'exactitude des coefficients, notamment du cosinus r et du sinus ρ , au regard des fonctions trigonométriques, de la relation de Pythagore ou de l'additivité de la variance :

- la moyenne $\bar{Y} = 4,875$
- $r^2 + \rho^2 = 1 = (0,873) + (0,127)$
- $\sigma_Y^2 = \rho^2 \sigma_Y^2 + r^2 \sigma_Y^2 = 18,109 = (2,299) + (15,810)$

3) La décomposition de la variance de Y

La décomposition de la variance de Y est ramenée à sa plus simple expression par la suppression du dénominateur commun n (nombre de couples, de lignes ou de facteurs constitutifs du vecteur) : $(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})$.

À savoir : *Variance totale = Variance résiduelle + Variance contrôlée.*

$(Y - \bar{Y})$	$(Y - Y')$	$(Y' - \bar{Y})$	<i>Discipline</i>
<i>Vt</i>	<i>Vr</i>	<i>Vc</i>	
1,125	-0,580	1,705	D1 : Droit
-2,875	0,073	-2,948	D2 : Sciences Eco
10,125	2,708	7,417	D3 : Lettres
1,125	-2,590	3,715	D4 : Sciences
-0,875	-0,994	0,119	D5 : Médecine-Dentaire

-3,875	-0,081	-3,794	D6 : Pharmacie
-3,875	1,717	-5,592	D7 : Pluridisciplinaire
-0,875	-0,254	-0,621	D8 : IUT
0,000	0,000	0,000	Total

Les sommes des « vecteurs gaussiens » sont nulles. Et donc les variables spectrales produites par la décomposition de la variance sont parfaitement indépendantes (suivant les théorèmes de Craig et de Cochran)

4) Les densités spectrales V_t , V_r et V_c

Les filtres transforment les vecteurs gaussiens en vecteurs linéaires dynamiques grâce à l'expression des densités factorielles. Ce sont les valeurs centrées réduites qui rendent compte des qualités inhérentes à la *liaison stochastique* des facteurs linéaires.

La lecture croisée (verticale et horizontale) des valeurs algébriques accentue la dynamique de la transformation linéaire, validée par le χ^2 de la distance dt .

Les valeurs limites des marges de distribution autour de la moyenne, fixées par les tables du χ^2 de Fisher, rendent toutes les valeurs immédiatement lisibles, en lecture verticale ou en lecture horizontale.

Limites de probabilités et normalité de distribution normale sont connues :

- a) de 95% pour un écart réduit z de ± 2 ($-1,96 \leq z \leq +1,96$)
- b) de 99% pour un écart réduit z de $\pm 2,58$ ($-2,58 \leq z \leq +2,58$)
- c) de sensiblement 100% pour un écart réduit z de ± 3

V_t	V_r	V_c	dt	<i>Discipline</i>
0,264	-0,383	0,429	0,361	D1 : Droit
-0,676	0,048	-0,741	0,219	D2 : Sciences Eco
2,379	1,786	1,865	0,710	D3 : Lettres
0,264	-1,708	0,934	1,364	D4 : Sciences
-0,206	-0,656	0,030	0,441	D5 : Médecine-Dentaire
-0,911	-0,053	-0,954	0,205	D6 : Pharmacie
-0,911	1,133	-1,406	1,103	D7 : Pluridisciplinaire
-0,206	-0,167	-0,156	0,071	D8 : IUT
0,000	0,000	0,000	4,473	

Les sommes des variables étant toujours nulles, les variables sont indépendantes.

La transformation de la variance σ_Y^2 de Y en fonction de X rend compte des qualités spectrales de la *liaison stochastique* pour chacun des facteurs et des vecteurs :

- a) si la valeur algébrique est *positive*, elle marque une *liaison*
- b) si la valeur algébrique est *négative*, elle marque une *résistance*

Il faut donc nécessairement tenir compte de la qualité linéaire du vecteur :

- a) le facteur Vt indique la *liaison globale* (estimation Y' de Y en X)
- b) le facteur Vr indique la *valeur résiduelle*, partie non expliquée de la variance
- c) le facteur Vc indique la *valeur contrôlée*, partie expliquée de la variance

C'est ainsi que 3 des 8 facteurs linéaires se distinguent immédiatement :

- a) **D3** avec un $dt = 0,710 > 0,500$ et un $Vt = 2,379 > +2$. Tous ces critères sont *positifs* : ce facteur est un élément de choix à la fois pour Y et pour X
- b) **D4** avec un $dt = 1,364 > 1$ et un $Vr = -1,708$: ce trait *négligatif* en fait un élément de rejet pour Y et donc de choix pour X
- c) **D7** avec un $dt = 1,103$ et un $Vr = + 1,103$: ce trait *positif* en fait un élément de choix pour Y et donc de rejet pour X

Comme les mesures des 5 autres facteurs sont « centrées » ou de plus faible intensité, ces éléments sont nécessairement liés. Les valeurs algébriques des écarts centrés et réduits fixent les qualités inhérentes aux densités spectrales et les limites de variation des nuages de points.

C'est ainsi que le tout se justifie par la partie et la partie par le tout. Le calcul algébrique est un calcul d'intégration (c'est à proprement parler la réduction des fractions à l'intégralité).

Tout est clairement mesuré et parfaitement identifié. C'est l'intégralité des *liaisons stochastiques* des vecteurs et des facteurs qui est offerte à *l'analyse qualitative* du corpus.

5) La distance quadratique dt du spectre de densité

Avec une valeur de 4,473 à 8 ddl, le χ^2 permet de savoir que l'ensemble des données est « contrôlé » à 81,2 %.

Dès lors toute l'analyse va s'efforcer de discerner les éléments qui perturbent cette distribution en vertu de la qualité des résidus. Tel est le rôle dévolu aux graphiques qui permettent de visualiser et d'identifier les phénomènes responsables des disparités qui se font jour.

Les valeurs de dt supérieures ou égales à 1 ($dt \geq 1$) fixent immédiatement l'attention (en rouge) sur les qualités particulières du vecteur linéaire, relevant d'une forte liaison ou, au contraire, d'une forte résistance. Mais toutes les valeurs expriment le moment d'inertie du spectre : elles sont parfaitement stables.

Dans ce cas, il faut avant regarder la valeur algébrique des résidus Vr :

- a) si la valeur de Vr est *positive*, il s'agit d'une *liaison*. Le facteur est un *élément préférentiel de Y*
- b) si la valeur de Vr est *négligative*, il s'agit d'une *résistance*. Le facteur est un *élément préférentiel de X*

C'est l'ensemble des valeurs algébriques qu'il faut prendre en compte pour projeter des *conclusions cohérentes et vérifiables*.

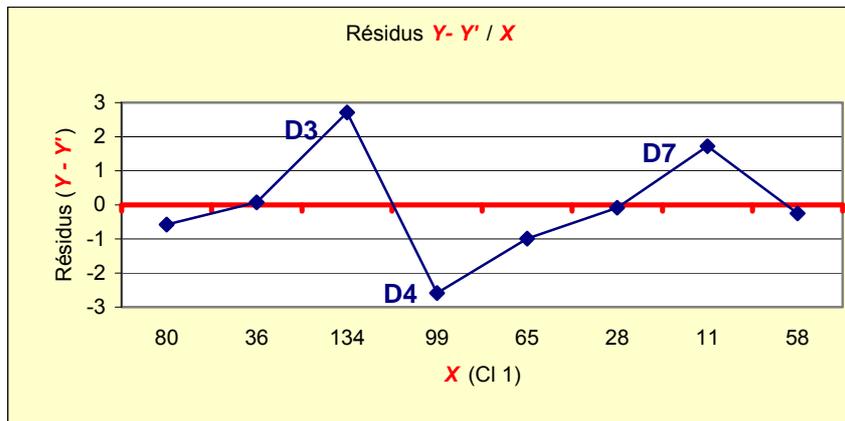
Les qualités de la distribution se reflètent dans les représentations graphiques où la partie s'intègre dans le tout et le tout coordonne les parties.

Les profils des liaisons stochastiques sont définis de façon dynamique et intégrale. On peut en vérifier la teneur qualitative.

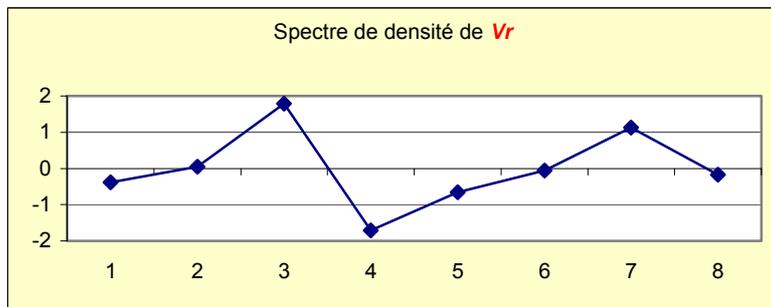
6) *Les graphiques des spectres de densité*

Regardons attentivement les différents graphiques à la lumière des échelles qui sont en marge, sans autre commentaire (ce qui serait présentement déplacé ou inapproprié).

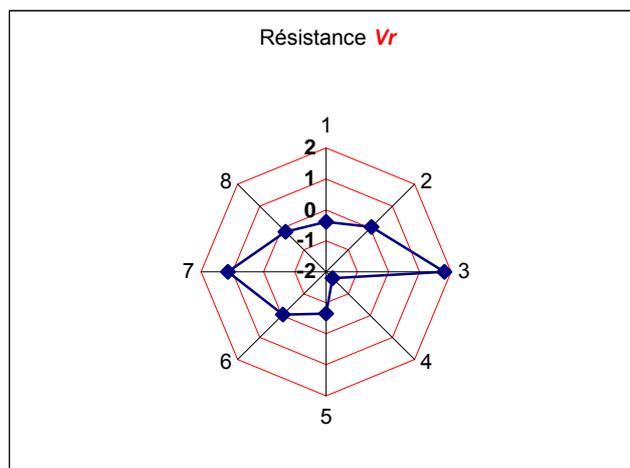
D'abord le graphique des résidus ($e = Y - Y'$) :



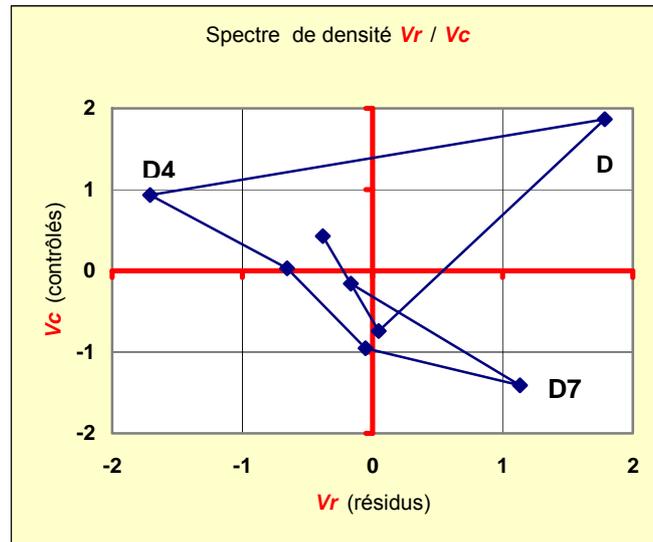
Comparons-le au spectre Vr des résidus centrés réduits :



Comparons-le maintenant au spectre en radar des résidus Vr :



Ajoutons-y en guise de commentaire le graphique du spectre de densité de V_r par rapport à V_c (des résidus par rapport aux éléments contrôlés) :

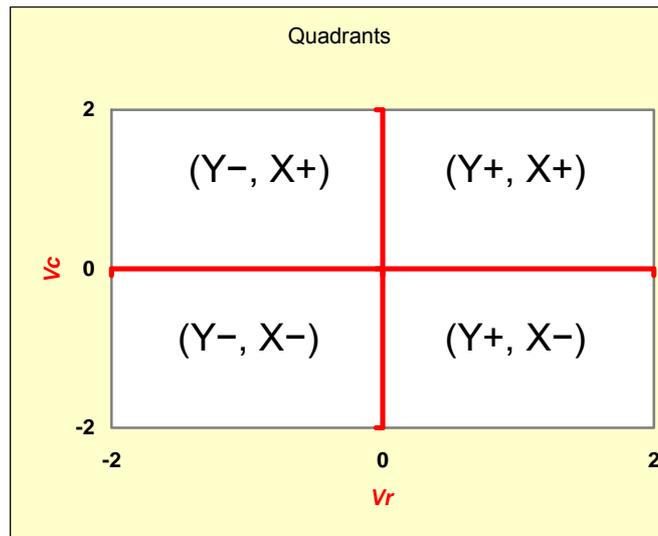


La synthèse est démonstrative.

L'analyse spectrale met en évidence les caractéristiques phénoménales de la distribution :

- 1) **D3**, dans le 1^{er} quadrant (++), montre que les 2 catégories sociales donnent sensiblement la même importance aux *Lettres*
- 2) **D7**, dans le 4^{ème} quadrant (+ -), montre en revanche une préférence marquée pour le *Pluridisciplinaire* par *Y* (C1 2)
- 3) **D4**, dans le 2^{ème} quadrant (- +), montre au contraire une préférence marquée pour les *Sciences* par *X* (C1 1).

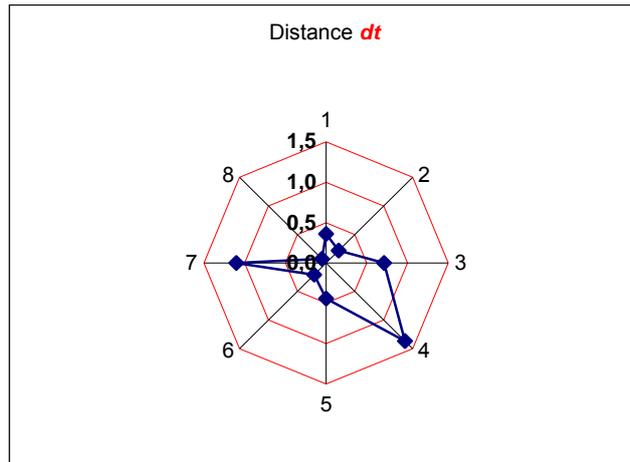
En fonction des coordonnées qui situent les points du nuage à l'intérieur des quadrants, la qualité de la *liaison stochastique* est immédiatement déchiffrée : la position du *critère Y* par rapport au *prédicteur X* dit la qualité de la liaison, de la résistances ou de l'opposition :



NB. Comme V_r est en abscisse et V_c en ordonnée, il faut lire le *critère* Y en abscisse et le *prédicteur* X en ordonnée. En cas de doute, revenir aux valeurs des primitives X et Y et de l'estimation Y' pour voir la qualité du *résidu* e ($e = Y - Y'$) :

- 1) Situés dans le 1^{er} quadrant, Y et X sont *positifs* : les variables sont *positivement liées*. Le facteur en cause est un *élément de choix* pour les deux variables.
- 2) Situés dans le 4^{ème} quadrant, Y est *positif* et X est *négatif* : les variables sont en opposition. Un V_r *positif* est la marque d'une surcharge de Y ou d'un déficit de X . Le facteur en cause est un *élément prépondérant de Y* , déficitaire en X .
- 3) Situés dans le 2^{ème} quadrant, au contraire, Y est *négatif* et X est *positif* : les variables sont dans une opposition inversée. Un V_r *négatif* est la marque d'un déficit de Y ou d'une surcharge de X . Le facteur en cause est un *élément prépondérant de X* , déficitaire en Y .
- 4) Situés dans le 3^{ème} quadrant, Y et X sont tous deux *négatifs* : les variables sont *négativement liées*. Le facteur en cause est un *élément de rejet* pour les deux variables.

Les *liaisons stochastiques* de Y' de Y en X sont explicitement confirmées par le radar des valeurs du χ^2 comme il ressort du graphique ci-après :



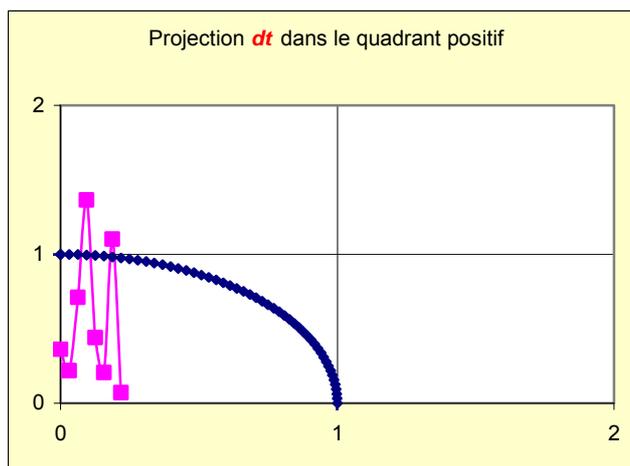
D4 et **D7** s'écartent du centre de gravité alors que **D3** s'en rapproche.

L'inertie du χ^2 (en radar) dit la spécificité de la densité spectrale.

Les spectres de densité fonctionnent comme des loupes grossissantes : ils mettent en évidence les caractéristiques fondamentales de la distribution, ils permettent de positionner les facteurs les uns par rapport aux autres et de les analyser avec perspicacité.

Alors la statistique remplit à merveille son rôle exploratoire et descriptif : elle « fait connaître » pour mieux « faire reconnaître ».

La projection orthogonale du χ^2 dans le 1^{er} quadrant du cercle carré confirme, si besoin est, la forte valeur résiduelle de **D4** et de **D7** (qui sont par ailleurs en opposition) :



Situés dans l'intervalle $+1 \leq dt \leq +2$, **D4** et **D7** sont excentrés : **D4** avec $dt = 1,364$ et **D7** avec $dt = 1,103$, deux distances dites « aberrantes », positivement significatives. Ce sont des valeurs remarquables.

Le χ^2 contribue pleinement à la valorisation de la description statistique en focalisant la stabilité spectrale de la discrimination.

Comme disait Paul Valéry : « Une théorie est d'autant plus 'scientifique' qu'elle fait entrevoir plus de *vérification* – qu'elle indique de résultats vérifiables. Une théorie n'est pas *vraie* ou non, elle est vérifiable ou non » (in *Cahiers II*. Paris : Gallimard, *La Pléiade*, p. 858).

5. Spectres de densité et Images de synthèse

La Macro, en ouvrant toutes grandes les portes de l'analyse spectrale, met à nu la fonction et la signification des coefficients dans l'étude de la régression linéaire et perce le secret des interférences vectorielles produites par le corpus tout entier.

La Macro permet de lire la Régression aussi bien dans *le sens direct de Y en X* que dans *le sens inverse de X en Y*.

Chemin faisant, nous observerons quelles sont la fonction et la signification des coefficients et des paramètres d'estimation, en tant que charges utiles de calcul et d'analyse de la *liaison stochastique*.

5.1 Les paramètres d'estimation, valeurs constantes et valeurs propres

Définition des paramètres d'estimation :

1) Paramètres d'estimation de Y' (de Y en X , sens direct) :

\bar{Y}	r	r^2	ρ	σ_Y	$u = \rho \cdot \sigma_Y$	$c = r \cdot \sigma_Y$
4,875	0,934	0,873	0,356	4,256	1,516	3,976

2) Paramètres d'estimation de X' (de X en Y , en sens inverse) :

\bar{X}	r	r^2	ρ	σ_X	$u = \rho \cdot \sigma_X$	$c = r \cdot \sigma_X$
63,875	0,934	0,873	0,356	37,595	13,395	35,127

Les valeurs de r , r^2 et ρ sont constantes, puisque ce sont les mêmes valeurs centrées du couple X/Y ou Y/X – $x = (X_i - \bar{X})$ et $y = (Y_i - \bar{Y})$ – qui entrent dans le calcul des coefficients.

Seules changent les valeurs propres de chaque variable : les moyennes \bar{X} et \bar{Y} , et les écarts quadratiques moyens u (de résistance) et c (de contrôle).

Variance et variabilité. Dans l'hypothèse d'une *liaison linéaire* entre X et Y , la connaissance de X permet d'expliquer :

- a) la *variance* σ_Y^2 de Y à concurrence de $r^2\%$
- b) la *variabilité* σ_Y de Y à concurrence de $(1 - \rho)\%$

En fonction des valeurs de r , on connaît systématiquement :

- a) $r^2\%$ le taux de *variance* σ_Y^2 de Y
- b) $(1 - \rho)\%$ le taux de *variabilité* σ_Y de Y

La *variabilité* de Y ne s'explique qu'à concurrence de 50 % avec un r de $\pm 0,866$ dans l'intervalle $0 \leq \theta \leq \sqrt{\frac{3}{2}}$ correspondant à un angle θ de 30° .

Il faut donc un $r \geq 0,866$ pour que la liaison stochastique soit significative positive :

Angle θ	r	ρ	r^2	$(1 - \rho)$
30°	0,866	0,500	0,75	0,500
45°	0,707	0,707	0,50	0,266
60°	0,500	0,866	0,25	0,134

Avec un $r = 0,934$ et $\rho = 0,356$, la *variabilité* de **CI 2** par rapport à celle de **CI 1** s'explique à concurrence de 64,4 % ($1 - 0,356 = 0,644$). Un pur problème de mécanique.

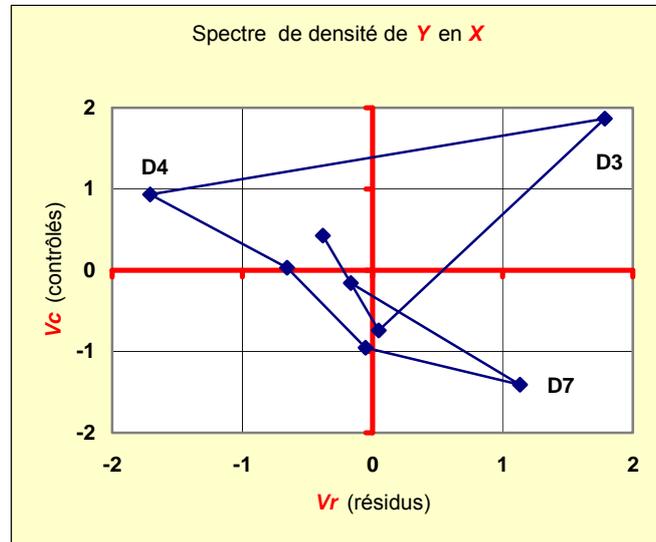
Théoriquement, la liaison entre **CI 1** et **CI 2** est de 65%. Ce qui signifie que le choix de ces deux classes sociales se fait de façon similaire sur 5 des 8 disciplines. Il faut alors tenir compte des 35% de résistance, soit 3 disciplines sur 8. Or nous avons vu que **D3**, **D4** et **D7** faisaient effectivement la différence. Une résistance d'ordre social.

3) *Fonction et signification des coefficients et des paramètres d'estimation*

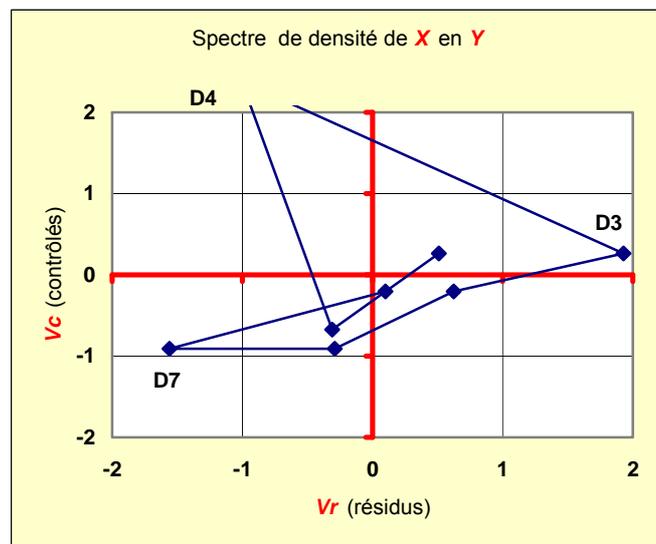
La Macro permet de déterminer directement les proportions d'éléments liés et d'éléments aberrants ou remarquables suivant la signification du coefficient de corrélation r .

- a) si le coefficient de corrélation $r \geq 0,866$, il indique une *hypothèse de liaison de Y expliquée à concurrence de $(1 - \rho)\%$ de n par la connaissance de X*. Les *éléments aberrants* sont dans une proportion complémentaire de $\rho\%$ de n .
- b) inversement, si le coefficient de corrélation $r < 0,866$, il indique une *hypothèse de non liaison*. Les *éléments aberrants de Y* sont *expliqués à concurrence de $(1 - \rho)\%$ de n par ceux de X*. Et les éléments de liaison sont dans la proportion complémentaire de $\rho\%$ de n .
- c) la Macro donne les proportions de liaison et d'aberration. Dans le cas présent, $n = 8$, $r = 0,934$ et $\rho = 0,356$. D'où *l'hypothèse de liaison* dans les proportions suivantes : la *liaison* ($1 - r = 0,644$) = 65 % de 8 = 5 éléments liés ; l'*aberration* ($r = 0,356$) = 35 % de 8 = 3 éléments « incontrôlés ».

- d) D'où le nuage de points déjà observé, de l'estimation Y' de Y en X , où l'on compte, sur les 8 éléments, 5 éléments de liaison et 3 éléments d'aberration (**D3**, **D4** et **D7**) :



- e) Comparons cette image à l'image du nuage de points de l'estimation inverse de X' de X en Y , où tous les éléments occupent des positions modifiées, et sensiblement pour 2 des 3 éléments d'aberration, **D4** et **D7** :



Les images de synthèse et les spectres de densité donnent la mesure exacte de la liaison stochastique. On a un éclairage global des relations factorielles intrinsèques au corpus, surtout si l'on compare les phénomènes dans des fenêtres de même taille.

Le calcul algébrique est un calcul d'intégration, de mesure et de comparaison. Le calcul arithmétique est un calcul de détermination et de contrôle. Et le calcul géométrique est un calcul de représentation et de visualisation.

5.2 Images spectrales et spectres de densité

Le croisement des variables entraîne le croisement des densités spectrales.

Les valeurs de V_t et de V_c qui sont celles de l'estimation Y' de Y en X , sont inversées dans l'estimation X' de X en Y de telle sorte que $V_t = V_c$ et $V_c = V_t$.

Seules changent les valeurs résiduelles de V_r qui sont propres à chaque estimation, et par conséquent les valeurs de dt .

D'où l'importance capitale des *résidus* tels qu'on peut les observer dans le croisement des variables, en passant de Y/X à X/Y :

1) Valeurs réduites du spectre de Y' (de Y en X)

V_t	V_r	V_c	dt	Discipline
0,264	-0,383	0,429	0,361	D1 : Droit
-0,676	0,048	-0,741	0,219	D2 : Sciences Eco
2,379	1,786	1,865	0,710	D3 : Lettres
0,264	-1,708	0,934	1,364	D4 : Sciences
-0,206	-0,656	0,030	0,441	D5 : Médecine-Dentaire
-0,911	-0,053	-0,954	0,205	D6 : Pharmacie
-0,911	1,133	-1,406	1,103	D7 : Pluridisciplinaire
-0,206	-0,167	-0,156	0,071	D8 : IUT
0,000	0,000	0,000	4,473	total

2) Valeurs réduites du spectre inversé de X' (de X en Y)

V_t	V_r	V_c	dt	Discipline
0,429	0,511	0,264	0,271	D1 : Droit
-0,741	-0,309	-0,676	0,034	D2 : Sciences Eco
1,865	-1,004	2,379	1,263	D3 : Lettres
0,934	1,929	0,264	1,208	D4 : Sciences
0,030	0,623	-0,206	0,464	D5 : Médecine-Dentaire
-0,954	-0,290	-0,911	0,038	D6 : Pharmacie
-1,406	-1,559	-0,911	0,801	D7 : Pluridisciplinaire
-0,156	0,101	-0,206	0,118	D8 : IUT
0,000	0,000	0,000	4,196	total

Les valeurs de V_t et de V_c sont effectivement croisées avec le croisement des variables : c' est une évidence. Laissons de côté les explications mathématiques.

En revanche, les valeurs de V_r et de dt sont propres à chaque cas.

Les valeurs algébriques de V_r sont ici quasiment opposées, dénotant une inversion de densité et d'intensité d'un tableau à l'autre, et donc un vecteur à l'autre.

Les spectres et les images de synthèse ouvrent la voie à l'analyse des phénomènes. Elles permettent de déceler ce qui n'était ni perceptible ni visible à l'œil nu. Elles permettent d'embrasser l'étendue des problèmes et d'envisager les réponses appropriées, le tout à la lumière de mesures constamment vérifiées et vérifiables.

On évitera néanmoins de les multiplier. On choisira parmi les plus représentatives, suivant la définition même de l'image qui est de faire voir, faire une synthèse de l'évidence.

6. Images et Moments d'inertie spectrale

Les notions de barycentre et de centre de gravité sont fondamentales en mécanique. Nous ne les perdons pas de vue dans la lecture des images de synthèse représentant les moments d'inertie de décomposition de la variance. Rappelons qu'il s'agit de variables qui suivent une loi normale LG (0 ; 1) : le barycentre est évident.

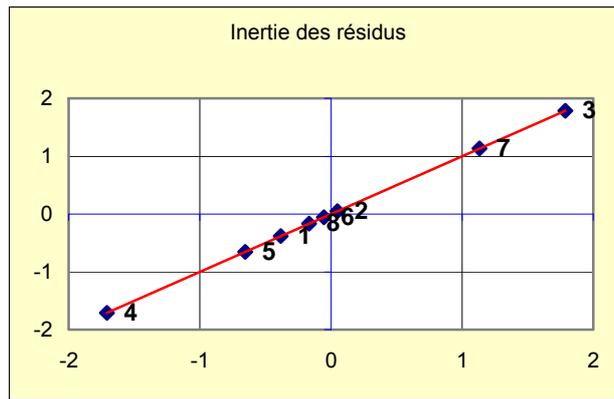
Parmi les graphiques possibles, nous en retiendrons deux :

- 1) la droite anamorphe de Henry, d'équation $y = x$, qui place les points du nuage sur une droite passant par l'origine des axes.
- 2) la parabole polynomiale d'équation $y = x^2$ qui place les points du nuage de façon symétrique de part et d'autre de l'axe des ordonnées, ce qui rend encore plus expressive l'image de la distribution.

6.1 La droite anamorphe des résidus

La solution la plus simple, c'est celle qui consiste à dédoubler la colonne des résidus pour tracer la droite anamorphe de Henry, d'équation $y = x$ (le barycentre est à l'origine des axes).

Considérons la droite anamorphe des résidus de Y en X , concernant les 2 premières variables du corpus :



Pour des raisons de clarté, nous désignons les points du nuage au moyen du seul nombre : nous mettons 4 au lieu de D4, 7 au lieu de D7 et 3 au lieu de D3, et ainsi de suite.

On retrouve ici les 5 éléments centrés ou liés et les 3 éléments aberrants. La hiérarchie est fixe et l'image est stable. La position des points rend compte de la qualité des facteurs.

6.2 La parabole polynomiale des résidus

La parabole exprime certainement de façon beaucoup plus sensible la qualité des facteurs, parce que le polynôme (d'équation du second degré $y = x^2$) positionne les points du nuage de façon à la fois stable et dynamique (ce qui est un paradoxe) de part et d'autre de l'axe des ordonnées. Le mouvement de la courbe exprime à la fois la convergence et la divergence par rapport au centre de gravité. Ce faisant, la parabole échappe à la platitude de la droite de Henry qui porte bien son nom d'anamorphose.

a) les données polynomiales

L'inertie totale du spectre est fixée par les carrés Vt^2, Vr^2, Vc^2 des densités Vt, Vr et Vc de décomposition et de transformation spectrale de la variance :

Prenons les carrés des densités du spectre d'estimation de Y (CI 2) en X (CI), les 2 premières variables du corpus :

Vt	Vr	Vc	Vt^2	Vr^2	Vc^2	<i>Discipline</i>
0,264	-0,383	0,429	0,070	0,147	0,184	D1 : Droit
-0,676	0,048	-0,741	0,456	0,002	0,550	D2 : Sciences Eco
2,379	1,786	1,865	5,661	3,190	3,479	D3 : Lettres
0,264	-1,708	0,934	0,070	2,918	0,873	D4 : Sciences
-0,206	-0,656	0,030	0,042	0,430	0,001	D5 : Médecine-Dentaire
-0,911	-0,053	-0,954	0,829	0,003	0,911	D6 : Pharmacie
-0,911	1,133	-1,406	0,829	1,283	1,978	D7 : Pluridisciplinaire
-0,206	-0,167	-0,156	0,042	0,028	0,024	D8 : IUT
0,000	0,000	0,000	8,000	8,000	8,000	<i>total</i>

Les sommes des densités sont nulles (*cf. supra* Chap. 3, 3) :

- 1) $\sum V_t = 0$
- 2) $\sum V_r = 0$
- 3) $\sum V_c = 0$

(résultats conformes aux théorèmes de Craig et de Cochran).

De même, pour des raisons mathématiques évidentes, les sommes des carrés des densités sont toutes égales à n . Le nombre n est ici égal à 8, le nombre de lignes, vecteurs ou facteurs appariés, la trace de la matrice transformée, et aussi le nombre de *ddl* des distances du χ^2 .

- 4) $\sum V_t^2 = n$
- 5) $\sum V_r^2 = n$
- 6) $\sum V_c^2 = n$

Le nombre n , c'est le *nombre* de points du nuage de la droite ou de la courbe représentant l'inertie totale de la matrice des données estimées et transformées (*Y' de Y en X*).

D'où la stabilité du polynôme (d'équation $y = x^2$) garantissant le moment d'inertie totale de la parabole symétrique.

Les valeurs algébriques de V_t , de V_r et de V_c , sont des valeurs d'intégration de la partie dans le tout. Ce ne sont pas à proprement parler des mesures puisqu'elles sont sans dimension, mais des valeurs de corrélation qui résultent de l'appariement des facteurs, sachant que la corrélation est intransitive.

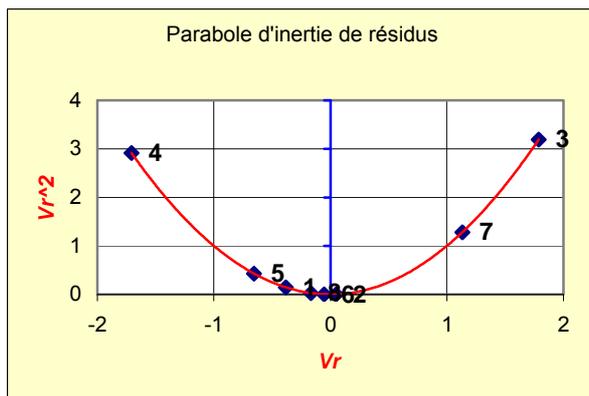
Sachant par ailleurs que les valeurs des vecteurs V_t et V_c sont inversées lorsqu'on inverse les variables appariées, il est aisé de travailler sur une seule matrice de données pour observer les appariements et en dégager les graphiques appropriés, sous forme de nuage de points ou de radars, par exemple.

b) la parabole polynomiale des résidus et images d'inertie

On peut aisément tracer trois paraboles complémentaires suivant les polynômes que l'on prend, puisqu'ils accompagnent la décomposition de la variance de Y en X :

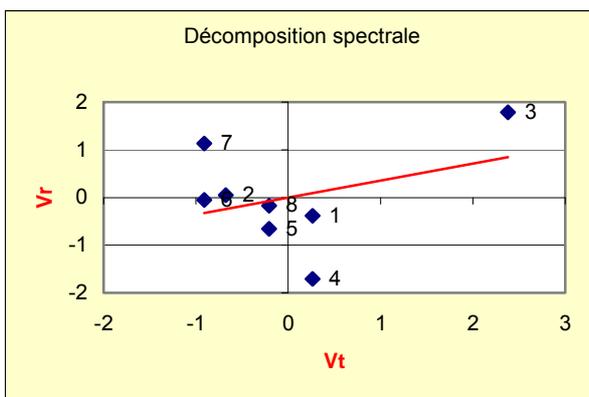
- 1) la parabole de la variance totale ayant V_t pour abscisse et V_t^2 pour ordonnée ;
- 2) la parabole de la partie contrôlée de la variance ayant V_c pour abscisse et V_c^2 pour ordonnée ;
- 3) la parabole des résidus du spectre ayant V_r pour abscisse et V_r^2 pour ordonnée.

C'est la parabole polynomiale des résidus qui nous intéresse au premier chef :

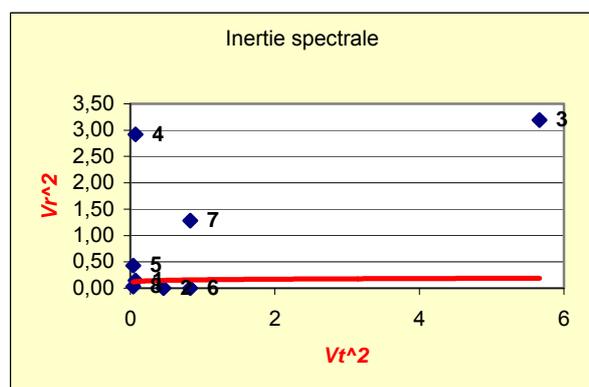


On voit immédiatement la position des points agglutinés (les liaisons) et la position des points excentrés en fonction de la hauteur qu'ils occupent sur l'une ou l'autre des deux branches de la parabole, comme dans le cas des points 3, 4 et 7. L'origine des axes est par excellence le barycentre du nuage de points.

Cette parabole est beaucoup plus démonstrative et beaucoup plus précise que ne peut être démonstratif et précis le nuage de points représentant l'inertie de la décomposition spectrale ayant V_t pour abscisse et V_r pour ordonnée :



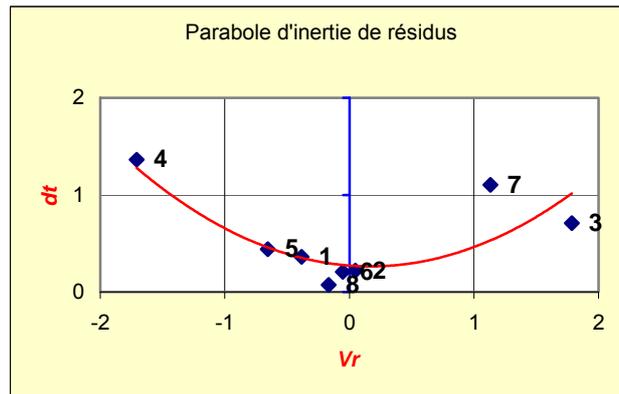
Ou encore l'image de l'inertie spectrale donnée par les carrés de V_r et de V_t :



On ira donc tout droit à la première parabole, la parabole polynomiale.

c) *la fausse parabole polynomiale produite par dt*

La fausse parabole polynomiale prend pour ordonnée non plus Vr^2 mais la distance dt (qui est elle-même un carré, un χ^2) :



Comme pour le χ^2 , les carrés des distances dt s'ajoutent et deviennent significatifs, alors que les valeurs algébriques se neutralisent.

L'appellation de fausse parabole est immédiatement justifiée puisque la courbe joue dans ce cas le rôle de la droite d'estimation le long de laquelle s'étalent les points du nuage.

La « vraie parabole d'inertie » donne une image fidèle de l'inertie spectrale. Elle permet d'embrasser d'un seul coup d'œil l'étendue des estimations et d'en saisir immédiatement les qualités et les intensités de recentrage ou de dispersion. Tous les éléments y sont parfaitement identifiés, elle ouvre la voie à une analyse fine et sûre. Le tout est éclairé par la partie et la partie par le tout.

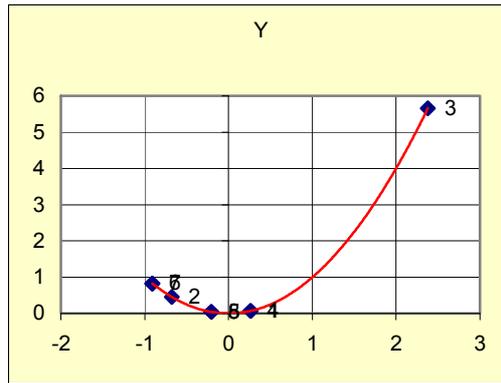
NB : La loi du khi-2 permet de définir le taux de probabilité de l'inertie spectrale dans chaque cas :

- 1) Dans la comparaison directe de Y en X elle est de 81,2 % ($p = 0,812$) pour un $\chi^2 = 4,473$ à 8 ddl
- 2) Dans la comparaison inversée de X en Y elle est de 83,9 % ($p = 0,839$) pour un $\chi^2 = 4,196$ à 8 ddl

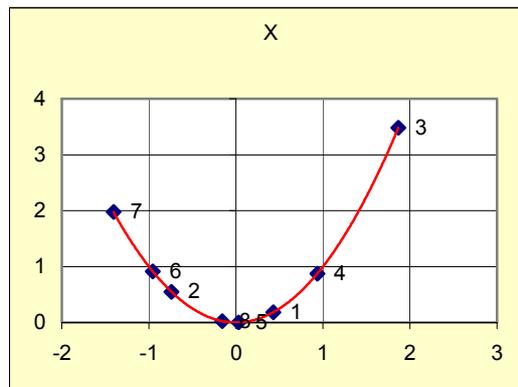
d) *les paraboles polynomiales de contrôle de la variance*

Juste à titre indicatif, voyons les deux paraboles de la variation totale Vt et de la variation contrôlée Vc concernant l'estimation Y' de Y en X des deux premières variables du corpus :

- a) parabole polynomiale de variation totale Vt qui caractérise Y :



b) parabole polynomiale de variation contrôlée V_c qui caractérise X :



Remarquons, juste en passant :

- 1) la position « privilégiée » de 3 (D3) sur la branche positive des paraboles de X et de Y , montrant que D3 est un facteur aussi important pour X que pour Y
- 2) la position « de rejet » de 7 (D7) sur la branche négative de la parabole de X
- 3) la position quasiment symétrique de 6 (D6) et de 4 (D4) sur la parabole de X

Les droites anamorphes de Henry sont beaucoup moins démonstratives, nous semble-t-il.

Les moments d'inertie doivent être pris en considération dans toute lecture d'une Table de Contingence (Voir Chapitre 4).

7. Conclusion

Le calcul d'une droite d'estimation doit toujours s'accompagner de celui du coefficient de corrélation qui précise la signification de *la liaison stochastique*, puisqu'il permet de formuler les hypothèses de travail pour entrer dans les arcanes de décomposition et de transformation spectrale de la variance afin de récolter les fruits d'une analyse pertinente débouchant sur des conclusions sûres, toujours vérifiées et toujours vérifiables.

L'Analyse Factorielle Discriminante (AFD) tire le meilleur parti des images de synthèse et des moments d'inertie délivrés par les densités spectrales des résidus qui jouent un rôle fondamental dans la comparaison, l'ajustement et l'estimation des variables appariées.

Le tout étant grandement facilité par le travail de la Macro qui montre à quel point la statistique, en jouant avec la loi du nombre, est un instrument de description et d'aide à l'interprétation, un instrument qui « fait connaître » pour mieux « faire reconnaître ».

Ajoutons enfin que la Macro permet de « filtrer automatiquement » les valeurs de chacune des colonnes de la page d'estimation pour mener à bien une analyse fine.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

4

Images et Mesures d’Inertie ou Procédures d’Analyse Factorielle Discriminante

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce quatrième chapitre, consacré à l’étude des Images et des Mesures d’Inertie, nous allons poursuivre et généraliser l’étude des images de synthèse décrivant l’inertie vectorielle entamée à la fin du chapitre précédent, concernant les procédures d’Analyse Discriminante qui sont au cœur de l’Analyse Statistique.

Nous allons procéder à la description de la structure des vecteurs analytiques en tournant les pages de la Macro qui effectue automatiquement l’Analyse Factorielle Discriminante (AFD).

Comme dans les chapitres précédents, nous utiliserons comme support technique les données du corpus présenté par G. Saporta à la page 151 des *Probabilités, Analyse des Données et Statistique*, Paris : Édit. Technip, 1990, tiré des *Données Sociales*, 3^e éd., INSEE, 1978.

1. Base des données et Matrice de calculs

La base des données n’est rien moins que la matrice des relevés concernant une population que l’on a recensée selon des critères prédéfinis.

La matrice des données est fondamentale : c’est la référence. Mais elle est immédiatement transformée en Table de Distribution des Fréquences (TDF), la première table de contingence que l’on va retrouver à la base de tous les calculs statistiques. Elle a de ce fait une importance capitale.

1.1 Base des données et TDF

La base des données n'est rien moins qu'une matrice des données qui se transforme instantanément en TDF dès lors que l'on remplace les données qualitatives marginales par les données numériques marginales.

1.1.1 Base des données (matrice initiale)

Discipline / CI	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
D1 : Droit	80	6	168	470	236	145	166	16	305
D2 : Sciences Eco	36	2	74	191	99	52	64	6	115
D3 : Lettres	134	15	312	806	493	281	401	27	624
D4 : Sciences	99	6	137	400	264	133	193	11	247
D5 : Médecine-Dentaire	65	4	208	876	281	135	127	8	301
D6 : Pharmacie	28	1	53	164	56	30	23	2	47
D7 : Pluridisciplinaire	11	1	21	45	36	20	28	2	42
D8 : IUT	58	4	62	79	87	54	129	8	90

Les classes sociales qui sont portées dans les colonnes et numérotées de 1 à 9 suivant la nomenclature CI 1, CI 2, ..., CI 9, constituent les variables du corpus :

CI 1	Exploitants agricoles
CI 2	Salariés agricoles
CI 3	Patrons
CI 4	Professions libérales et Cadres supérieurs
CI 5	Cadres moyens
CI 6	Employés
CI 7	Ouvriers
CI 8	Personnel de service
CI 9	Autres

1.1.2 La Table de Distribution des Fréquences (TDF)

La Table de Distribution des Fréquences (TDF) reprend les valeurs de la base des données en remplaçant les valeurs nominales par les valeurs marginales qui vont servir aux calculs de probabilités (Voir Chap. 1, 1)

D	E	F	G	H	I	J	K	L	M
Fréquences	CI 1	CI 2	CI 3	CI 4	CI 5	CI 6	CI 7	CI 8	CI 9
1592	80	6	168	470	236	145	166	16	305
639	36	2	74	191	99	52	64	6	115
3093	134	15	312	806	493	281	401	27	624
1490	99	6	137	400	264	133	193	11	247
2005	65	4	208	876	281	135	127	8	301
404	28	1	53	164	56	30	23	2	47

206	11	1	21	45	36	20	28	2	42
571	58	4	62	79	87	54	129	8	90
10000	511	39	1035	3031	1552	850	1131	80	1771

Les valeurs marginales permettent de calculer les probabilités « p » et les probabilités contraires « q » de chaque variable aléatoire.

La somme des « p » vaut 1 et la somme des « q » vaut 8 ($9 - 1$), ou, plus généralement : $\sum p = 1$ et $\sum q = (n - 1)$, n étant le nombre de variables.

	D	E	F	G	H	I	J	K	L	M
1	10000	511	39	1035	3031	1552	850	1131	80	1771
2	p	0,0511	0,0039	0,1035	0,3031	0,1552	0,085	0,1131	0,008	0,1771
3	q	0,9489	0,9961	0,8965	0,6969	0,8448	0,915	0,8869	0,992	0,8229

Les probabilités « p » et « q » sont des paramètres rigides, et tellement rigides qu'ils conditionnent tous les calculs statistiques qui s'ensuivent.

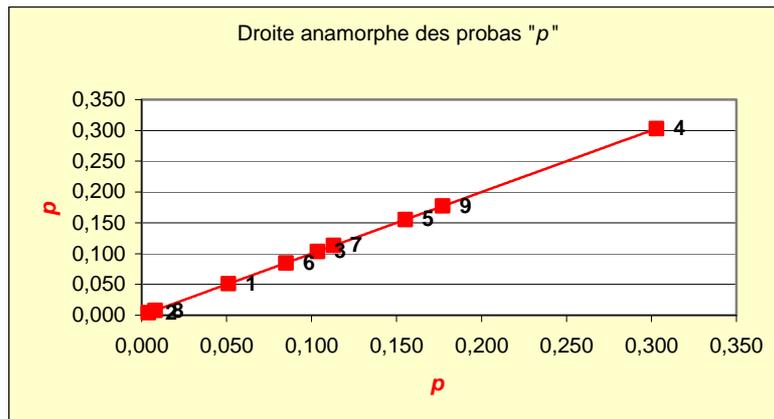
1.2 Les mesures rigides de la Table de Distribution des Fréquences

La probabilité « p » donne le tempo de la mesure statistique en rythmant la métrique des calculs des densités, des corrélations, des estimations et des discriminations factorielles. Elle est fondamentale. La probabilité contraire « q » n'est qu'une probabilité complémentaire.

La probabilité « p » est *la mesure rigide* par excellence. Elle permet de focaliser et de visualiser la distribution (hiérarchique) des variables au moyen d'une équation affine $y = x$ qui produit la droite de Henry (Voir Chap. 3, 6.2) ou, mieux encore, d'une équation du second degré $y = x^2$ qui produit la parabole polynomiale permettant de visualiser la position symétrique des facteurs d'abscisse positive ou négative sur l'une ou sur l'autre branche de part et d'autre de l'axe des ordonnées.

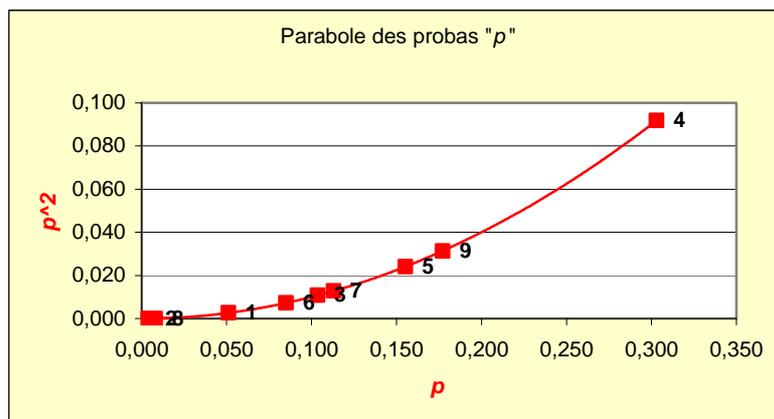
1.2.1 la droite anamorphe de Henry

La droite anamorphe de Henry donne l'image du nuage de points situé sur une droite d'équation $y = x$. On prend les mêmes valeurs en abscisse et en ordonnée. La stabilité est parfaite. La hiérarchie des variables est parfaitement visualisée.



1.2.2 la parabole polynomiale

La parabole polynomiale présente un avantage certain sur la droite anamorphe, même si la stabilité des positions est rigoureusement la même. De part sa forme, la courbe de la parabole exprime en apparence une dynamique de ce qui n'est en réalité que stabilité et fixité. Ce faisant, elle rend le graphique beaucoup plus sensible et beaucoup plus expressif.



Cette image est primordiale et fondamentale. Elle prend en considération les données primordiales et fondamentales du calcul statistique, celles de la probabilité. C'est l'image la plus représentative du nuage de points, puisqu'elle les présente sous leur meilleur jour. C'est d'ailleurs le but de la géométrie que de « faire voir à l'esprit ».

On pourrait renouveler ces opérations avec les valeurs de la probabilité complémentaire « q ». Les résultats seraient inversés, comme sont inversés le *cosinus* et le *sinus* dans la métrique R (Voir Chap. 2, 2 et 5, *la métrique R*).

On pourrait encore prendre « p » en abscisse et « q » en ordonnée, la stabilité serait certes garantie, mais il n'est pas sûr que la projection de la hiérarchie puisse se faire sous un meilleur jour.

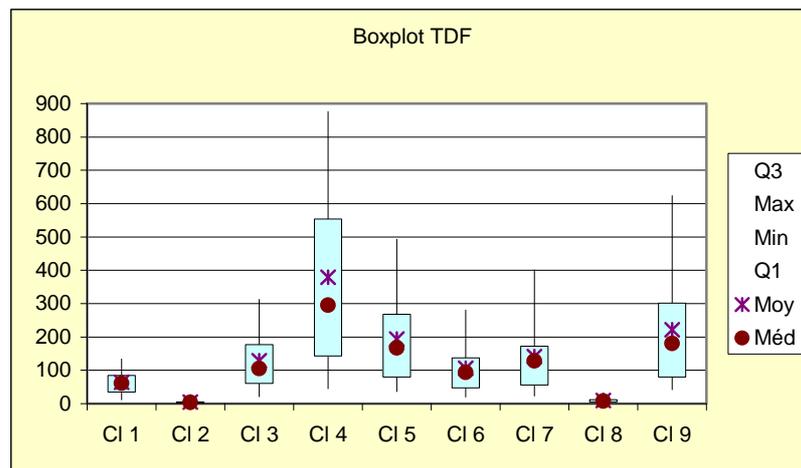
N'oublions pas l'adage qui dit que les mathématiques c'est l'art de raisonner juste sur des figures fausses. Si ce but est ici atteint, on n'a pas tout perdu.

1.3 Images et mesures d'inertie de la TDF

Parmi les images les plus en vue, on peut considérer que la méthode du *boxplot* proposée par Tukey est la plus connue. Nous allons de ce fait directement aux résultats en rappelant la méthode des calculs effectués pour voir que l'image obtenue est sans doute attrayante, mais dans la pratique peu convaincante ou peu probante.

1.3.1 la *boxplot* de la TDF

En utilisant les fonctions « quartiles » prédéfinies d'Excel on calcule dans l'ordre Q3, Max, Min, Q1 qui donnent l'image de base du « graphique boursier » que l'on complète en faisant glisser successivement le nuage de points de la Moyenne, de la Médiane (Q2) et les titres des variables. Rien de plus simple.

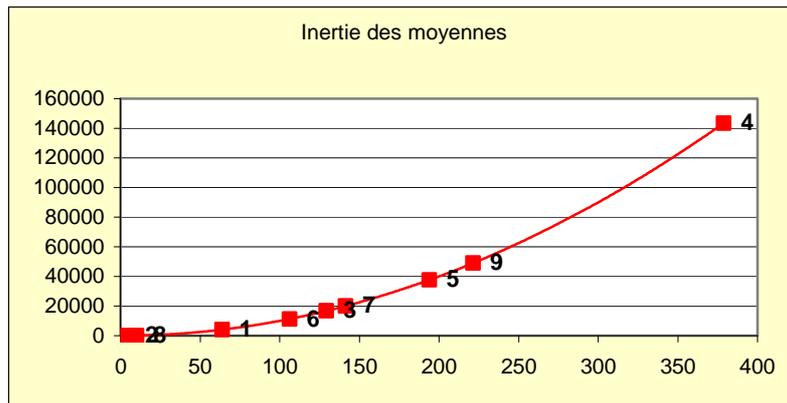


Le graphique est certes chatoyant, mais peu lisible. On voit que CI 4 occupe un espace imposant contrairement à CI 2 ou CI 8. On voit bien qu'il y a des différences, mais on ne peut rien dire de conséquent.

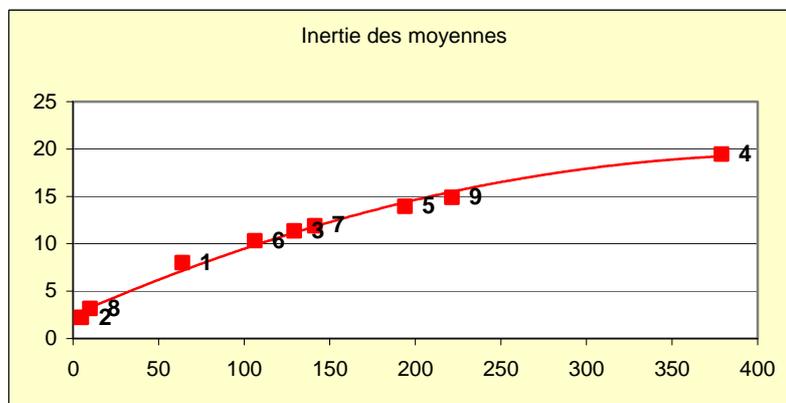
1.3.2 les paraboles polynomiales de la TDF

Etant donné que l'inertie est la moyenne pondérée des carrés des points du nuage au centre de gravité, on pourrait multiplier les graphiques à partir des carrés ou des racines carrés de la moyenne, de la médiane, de la différence quartale, etc. Elles doivent toutes refléter les qualités fondamentales de la distribution déjà données par les « probas », bien que présentées sous des angles différents à cause des rotations inhérentes aux méthodes de calcul.

Parmi les paraboles polynomiales les plus proches de la réalité descriptive des calculs de base, il faut compter la parabole d'inertie des carrés des moyennes d'équation $y = x^2$:



Ou encore la parabole d'inertie des racines des moyennes d'équation $y = \sqrt{x}$:



Les distances (hiérarchiques) des points du nuage sont identiques.

On remarque que la parabole des carrés des moyennes est identique à la parabole des « probas », et donc semblable à la parabole des racines des moyennes.

Quoi qu'il en soit, on a une « image vectorielle » de la TDF, équivalente de l'ACP (Analyse en Composantes Principales), mais on n'est pas encore entré dans l'AFD, (Analyse Factorielle Discriminante) à proprement parler, comme on peut le faire avec la TDR (Table des densités ou des écarts centrés réduits). (Voir Chap. 1)

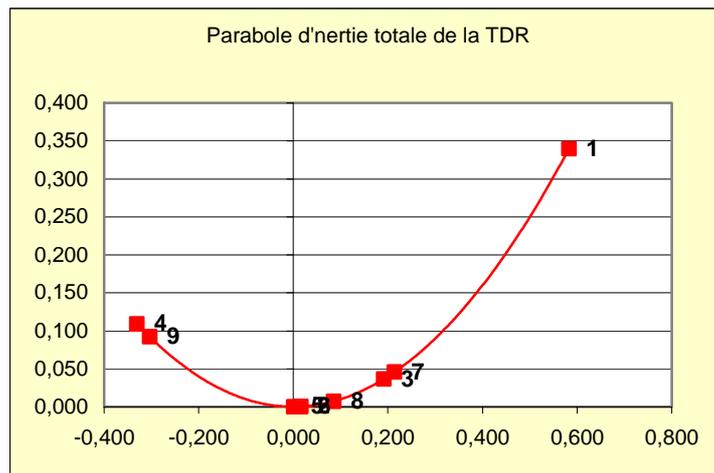
2. Le pouvoir discriminant de la TDR

Comme la TDR (table des écarts centrés réduits) a été présentée au chapitre 1, nous allons droit au but. Elle figure à la page 3 de la Macro.

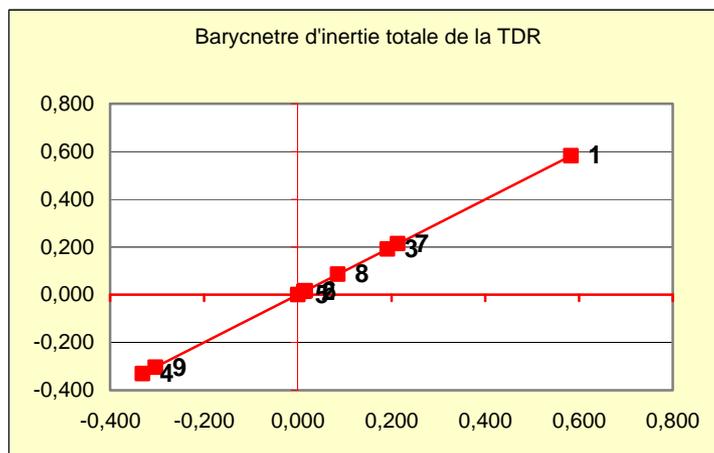
2.1 La discrimination globale

Les moyennes des écarts réduits et les carrés (pour le calcul du khi2) donnés en tête de la table, permettent de dégager immédiatement l'image de synthèse de la densité globale ou de l'inertie totale du corpus.

2.1 la parabole polynomiale de densité de la TDR



2.2 la droite anamorphe de l'inertie totale de la TDR



Tout commentaire est ici superflu.

2.2 La discrimination factorielle

La discrimination factorielle est par essence l'affaire de la TDR où tout n'est que valeurs algébriques intégrant la partie dans le tout, selon la définition même du « calcul algébrique » :

la partie qui se projette dans le tout et le tout qui se reflète dans la partie. L’algèbre est le calcul par excellence de la comparaison et de la description.

Alors, quel que soit le sens de lecture de la TDR, horizontale ou verticale, on a toujours affaire à des vecteurs inertes où les nuages de points sont fixes : tous les points se répartissent autour de la droite de moyenne réduite à 0 (zéro), suivant la loi normale LG (0 ; 1).

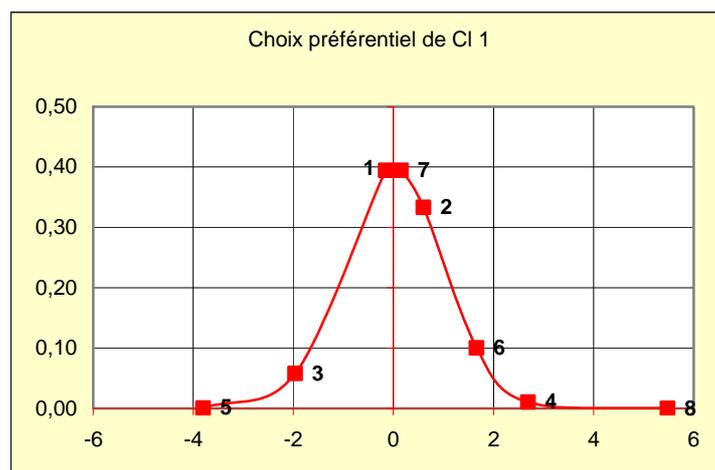
Quelle qu’en soit la nature, toute image est ici une image de synthèse, vectorielle ou factorielle, une image discriminante.

Recourir ici à la *boxplot* de Tukey où la moyenne est réduite à 0 (zéro), n’aurait aucun sens. Ce serait d’autant plus absurde que la TDR donne les densités des composantes : la valeur centrée réduite est la mesure comparative par excellence. Mais, les limites de calcul proposées par Tukey pour l’établissement de la *boxplot* (1,5dF ou différence interquartiles) sont moins précises que celles fournies par la TDR. En effet, la valeurs de Tukey *tendent grosso modo vers 2,7 de la valeur centrée réduite*. Or, la TDR donne les valeurs exactes de toutes les densités de la matrice.

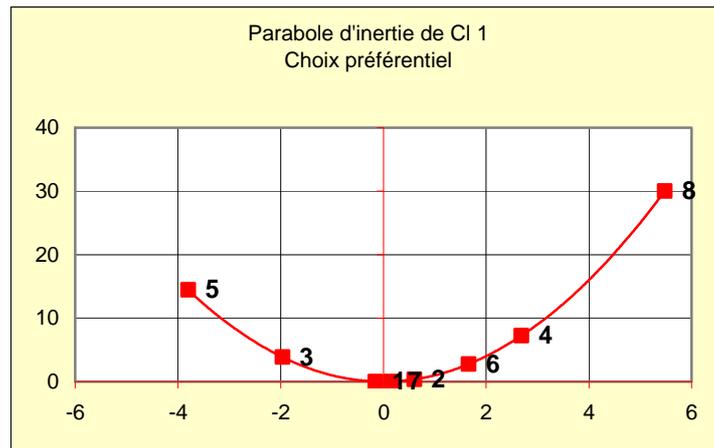
En revanche, on peut aussi bien tracer la droite anamorphe que la parabole polynomiale pour toute ligne ou pour toute colonne de la TDR. Elles ont toutes deux un sens et une signification.

2.3 La courbe en cloche et la parabole d’inertie

On peut tracer la courbe en cloche de Laplace-Gauss pour chacun des vecteurs ou pour chacune des variables, comme ici concernant les choix préférentiels des disciplines étudiées par le groupe CI 1 (des fils d’*Exploitants agricoles*) :



Cette courbe (sous forme de parabole renversée) en dit long sur la fonction discriminante de la TDR. (Pour tout ce qui est commentaires de méthode, voir Chap. 1).



La comparaison entre la cloche et la parabole est immédiate. Ce n'est pas la méthode de calcul des inerties qui va changer la nature du problème. La parabole est aussi démonstrative que la cloche, en outre elle a l'avantage d'être immédiatement accessible.

3. La puissance discriminante de la règle

La cinquième page de la Macro comporte un outil performant pour ce qui est de la discrimination et de la lemmatisation.

La *discrimination*, *stricto sensu*, c'est la réduction à l'unité des composantes factorielles.

Et, à l'inverse, la *lemmatisation*, c'est, *stricto sensu*, la formation d'un vecteur par un assemblage factoriel. Selon le type de corpus analysé, cet assemblage est fonction des paramètres du corpus :

1. s'il s'agit d'un corpus d'ordre médical : regroupement de traitements, de résultats d'analyses sanguines, d'expérimentations diverses...
2. s'il s'agit d'un corpus d'ordre économique : regroupement de relevés catégoriels, de performances techniques ou marchandes...
3. s'il s'agit d'un corpus d'ordre financier : regroupement de recettes, de dépenses et de bilans...
4. s'il s'agit d'un corpus de nature littéraire : regroupement de termes par affinité thématique, sémantique, synonymique, grammaticale...
5. s'il s'agit d'un corpus de nature écologique : regroupement des mesures chimiques, des mesures de pollution, des mesures d'amélioration dans le temps et dans l'espace...
6. s'il s'agit d'un suivi scolaire : regroupement en fonction des notes, des matières ou des disciplines, des classes, des années, des établissements, des régions...

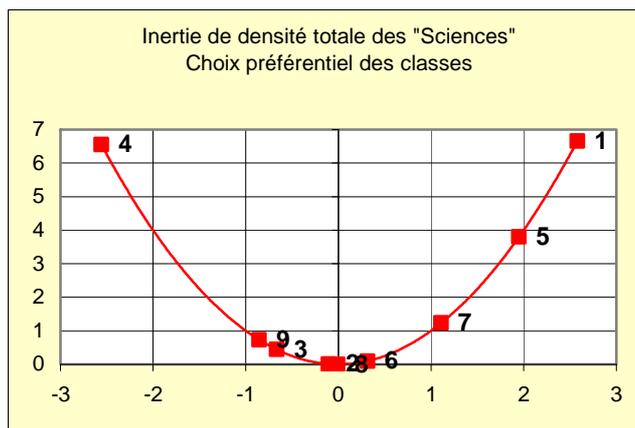
On peut tout observer de ce qui constitue une matrice de relevés statistiques.

À titre d'exemple, par *lemmatisation* on pourrait regrouper 'artificiellement' :

1. dans le domaine des disciplines : les sciences et techniques (D2, D4 et D8), les disciplines de la santé, médecine, dentaire et pharmacie (D5 et D6)...

2. dans le domaine des origines sociales : les fils d'agriculteurs (Cl 1 et Cl 2), les fils des cadres supérieurs et moyens (Cl 3, Cl 4 et Cl 5), les fils de la classe ouvrière (Cl 6 et Cl 7) et le «restant» (Cl 7 et Cl 8).

Pour ce faire, on forme un seul vecteur par addition des données factorielles qui va donner lieu à un nuage de points d'équation $y = x^2$. Et la parabole est à *proprement parler* une image de synthèse, qui est identique à celles que nous connaissons parfaitement :



Les « Sciences » (D2 et D4) attirent les classes 1 et 5 (Cl 1 et Cl 5), les fils d'exploitants agricoles et de cadres moyens, et sont rejetées par la classe 4 (Cl 4), les fils des professions libérales et des cadres supérieurs. Sans autre commentaire.

La *lemmatisation* et la *discrimination* vont affiner l'analyse des (données produites par la matrice des) relevés statistiques. (Voir Chap. 5)

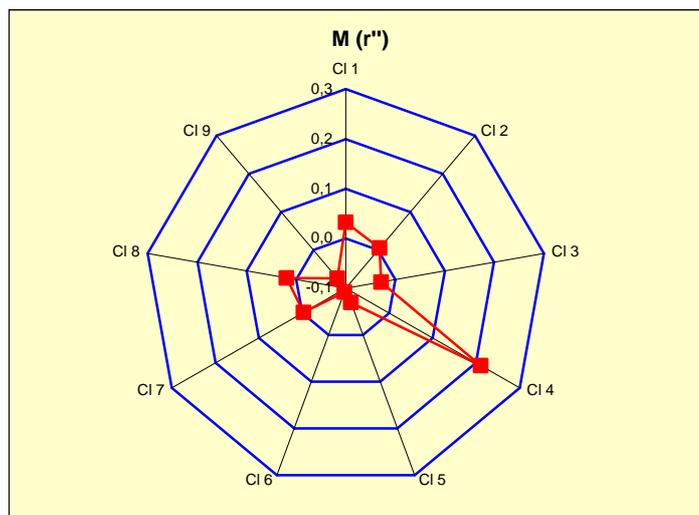
4. La table de corrélation et la métrique R

La table de corrélation et la métrique R ont partie liée. Elles produisent systématiquement tous les graphiques utiles à la mesure des inerties vectorielles et factorielles et à l'observation des projections obliques ou orthogonales suivant des angles de rotation propres à présenter les phénomènes sous leur meilleur jour, pour en « faire voir » les caractéristiques.

Les pages 6 et 7 de la Macro ont été longuement décrites, nous n'y reviendrons pas. Nous laissons au lecteur d'en découvrir toute la puissance qui est tout à la fois une puissance de calcul et de représentation graphique. Toutes les images de synthèse que l'on en retire sont des images d'inertie totale, qui rendent parfaitement compte des liaisons vectorielles et factorielles concernant les qualités phénoménales du corpus analysé.

La métrique R opère ici le croisement de l'ACP et de l'AFD, alliant lecture verticale et lecture horizontale des données de base.

À titre d'exemple, il suffit de regarder le radar des *sinus* concernant le comportement des 9 classes sociales face aux études proposées :



Le comportement des enfants des Professions libérales et des Cadres supérieurs, la classe Cl 4 du corpus, est pour le moins étrange : il se singularise par un « rejet » global des 8 disciplines retenues. Or, nous le savons, ce rejet est motivé par le choix ou par l'attrait hautement préférentiel de la discipline de la « santé », médecine et dentaire, D5.

Les images de synthèse méritent parfaitement leur nom dans la mesure où elles font la synthèse des résultats du parcours analytique qui a commencé avec les « probas » de la TDF, lesquels trouvent ici un premier aboutissement synthétique, sous forme d'ACP.

On peut rapprocher ce « radar » (des *sinus*) des paraboles polynomiales (des « probas » ou des moyennes). Il n'y a pas d'antinomie. Au contraire, l'un éclaire l'autre. (Voir *supra* Chap. 4, 1)

Bien entendu, la *box-plot* de Tukey n'a aucun sens ni aucune raison d'être ici non plus. Que signifierait d'ailleurs un graphique qui limiterait les « moustaches » à 1, la valeur plafond de la corrélation linéaire ?

5. L'estimation ou l'AFD par excellence

Nous ne reviendrons pas ici sur les images de synthèse et sur les calculs d'inertie concernant le travail réalisé à la page 8 de la Macro. (Voir Chap. 3)

Pourtant, comme on peut le voir dans le chapitre suivant, le chapitre 5, il y a beaucoup à attendre des capacités de calcul de la Macro en matière d'analyse discriminante pour le traitement d'un corpus complexe et multidimensionnel, notamment de la page 8 d'*Estimation*.

Nous allons voir que l'AFD dispose d'une panoplie de filtres qui peuvent pousser l'analyse statistique dans ses derniers retranchements, par le biais de la Macro qui joue un rôle essentiel et éminent. C'est ainsi que l'*Analyse Factorielle Discriminante* prend tout son sens.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

5

Traitement d'un corpus complexe ou multidimensionnel

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce cinquième chapitre, nous allons aborder le traitement statistique d'un corpus complexe ou multidimensionnel.

Le corpus complexe ou multidimensionnel s'oppose au corpus simple. Par corpus simple il faut entendre un corpus dont la dimension linéaire ou vectorielle est qualitativement réduite à l'unité, comme dans le corpus emprunté à G. Saporta et à l'INSEE. C'est un corpus à lecture variable comme on l'a vu, puisque nous avons mis les classes sociales dans les colonnes (les variables allant de 1 à 9, de CI 1 à CI 9) et les disciplines dans les lignes (les vecteurs étant formé de 8 facteurs qualitatifs simples, les disciplines classées de 1 à 8, de D1 à D8). Par corpus complexe ou multidimensionnel il faut, en revanche, entendre un corpus dont la dimension linéaire est complexe, puisqu'elle peut être formée de vecteurs simples ou de vecteurs composites, comme dans le corpus que nous allons prendre comme support.

Tout l'intérêt de ce chapitre consiste dans l'Analyse Factorielle Discriminante que la Macro va en faire. Mais auparavant faut-il « fabriquer le corpus statistique », c'est-à-dire faire le recensement et le classement des éléments que l'on veut analyser.

Pour ce faire, nous allons prendre comme support les 8 contes en prose de Charles Perrault, tirés des *Contes de ma mère l'Oye*. Comment recenser la population « verbale » qui les compose ? Comment élaborer les matrices de la base des données ? Comment faire les tables de contingence, la TDF (table de distribution des fréquences) et la TDR (table des écarts réduits), ou encore la table de corrélation, faire une lemmatisation ou une discrimination, faire une ACP (analyse en composantes principales) ou une AFD (analyse factorielle discriminante), une estimation ? Comment faire les graphiques appropriés aux différents paramètres ? Bref, comment mener à bien une analyse statistique de ce type ?

Le corpus littéraire des 8 contes de Perrault a été traité par STABLEX, un logiciel ayant pour auteurs André CAMLONG et Thierry BELTRAN, édité par la *Pirus Tecnologia* au Brésil (São Paulo), et disponible sur le marché en France chez () ou au Brésil chez pedropaulo@pirus.com.br ou www.pirus.com.br (Tel. : 00 55 11 6854.8593).

La présentation de la matrice des données lexicales et des tables de contingence va nous permettre de comprendre d'entrée de jeu ce que c'est qu'un corpus complexe ou multidimensionnel et ensuite d'entrer dans les arcanes de l'analyse statistique descriptive, objective et inductive.

C'est *la statistique à la portée de tous*.

La base des données, c'est la matrice des relevés contenant la population lexicale des 8 contes de Perrault classés de T1 à T8 suivant l'ordre chronologique des textes.

Le dépouillement lexical effectué par STABLEX est immédiatement pris en charge par la Macro qui n'a guère plus de secrets pour nous maintenant.

Nous allons souligner les nouveautés du traitement effectué dans les 7 premières pages de la Macro pour nous attarder sur l'*AFD* et l'estimation de la 8^{ème} page :

- 1) le Lexique, matrice et base des données lexicales
- 2) la TDF ou table de distribution des fréquences
- 3) la TDR ou table des écarts réduits
- 4) les graphiques de base
- 5) la règle ou outil de lemmatisation et de discrimination
- 6) la matrice de corrélation
- 7) l'*ACP* et la métrique R
- 8) l'*AFD* et l'estimation

1. Base des données et tables de contingence

Les 5 premières pages de la Macro sont d'un seul tenant. Le travail statistique y est effectué d'un seul trait.

1.1 Le Lexique

Il suffit d'examiner la matrice du Lexique pour comprendre ce que c'est qu'un corpus complexe ou multidimensionnel.

Le Lexique, en tant que matrice, comprend 10 colonnes et 2698 lignes.

Dans la première colonne, les « *Mots* différents ». Dans la 2^{ème} colonne, les *Occurrences* rangées par ordre décroissant, elles orientent la matrice. Dans les 8 colonnes suivantes, la *distribution des fréquences* (emploi des mots) dans chacun des 8 contes, rangés de T1 à T8.

C'est le *status de la population recensée*.

Observons le début de cette matrice :

Mot	Occ	T1	T2	T3	T4	T5	T6	T7	T8
de	615	114	25	71	67	28	80	88	142
la	594	158	30	75	25	44	84	87	91
et	533	115	23	68	48	27	71	68	113
le	450	80	31	38	103	11	38	56	93
il	376	85	11	33	45	10	28	38	126
qu'	345	83	6	31	29	16	52	65	63
à	331	73	11	38	38	20	42	36	73
que	320	54	15	35	35	13	36	69	63
elle	290	67	7	37	1	21	75	58	24
en	256	54	13	21	22	11	37	49	49
les	249	54	6	36	21	6	43	12	71
qui	236	32	16	20	26	16	31	42	53

Au fur et à mesure que l'on descend dans l'échelle des *Occurrences*, le nombre de *mots* ou *vocables* augmente à chaque fréquence pour atteindre des quantités importantes à la fréquence 1, les *hapax* ou mots qui ne sont utilisés qu'une seule fois dans l'un des 8 contes. On va le percevoir très nettement en regardant la TDF.

Comme on le voit, la *Base des données* est, à une nuance près, sensiblement identique à celle des données de G. Saporta, in Chap. 1, 1 : les items sont classés suivant l'ordre décroissant des Occurrences, qui constituent la colonne des données marginales qui figure dans la TDF, in Chap. 1,2.

De telle sorte que la matrice du *Lexique* préfigure la TDF.

1.2 La TDF

La complexité de la matrice du *Lexique* conditionne la complexité de la TDF (table de distribution des fréquences).

Voyons-en les 10 premières et les 10 dernières lignes :

Rang	Occ	Nbre	Fréq	T1	T2	T3	T4	T5	T6	T7	T8
1	615	1	615	114	25	71	67	28	80	88	142
2	594	1	594	158	30	75	25	44	84	87	91
3	533	1	533	115	23	68	48	27	71	68	113
4	450	1	450	80	31	38	103	11	38	56	93
5	376	1	376	85	11	33	45	10	28	38	126
6	345	1	345	83	6	31	29	16	52	65	63
7	331	1	331	73	11	38	38	20	42	36	73
8	320	1	320	54	15	35	35	13	36	69	63
9	290	1	290	67	7	37	1	21	75	58	24
10	256	1	256	54	13	21	22	11	37	49	49
83	200	20	10	29	9	28	13	14	26	38	43
84	252	28	9	49	12	43	18	10	30	29	61
85	232	29	8	37	4	36	17	18	39	30	51

86	238	34	7	46	21	24	13	21	20	40	53
87	312	52	6	57	11	38	35	17	26	41	87
88	405	81	5	93	34	48	24	17	63	47	79
89	432	108	4	96	25	44	38	25	70	62	72
90	552	184	3	103	12	65	67	33	76	93	103
91	862	431	2	170	31	116	89	43	126	112	175
92	1495	1495	1	342	44	160	144	62	198	221	324

Colonne 1, les *Rangs* de 1 à 92.

Colonne 2, les *Occurrences* par ordre décroissant.

Colonne 3, le *Nombre de Mots ou vocables* de même Fréquence dans la ligne.

Colonne 4, les *Fréquences*.

Et colonnes 5 à 12, la *distribution des fréquences* dans les 8 textes, classés de T1 à T8.

Commentaire rapide :

- 1) Au rang 83, il y a 200 occurrences venant de 20 mots ou vocables de fréquence 10.
- 2) Au rang 92, il y a 1495 occurrences venant de 1495 mots de fréquence 1.
- 3) Au rang 91, il y a 862 occurrences venant de 431 mots de fréquence 2.

On connaît la distribution des fréquences dans les 8 textes, les 8 colonnes, les 8 variables.

De haut en bas de la table, on trouve d'abord les mots de haute fréquence, pour la plupart des mots grammaticaux de fréquence élevée. Puis, au fur et à mesure qu'on descend, on trouve plusieurs mots de même fréquence à la même ligne, des « mots notionnels » en règle générale, jusqu'aux mots de fréquence 1, les hapax, à la dernière ligne, la 92^{ème}.

En tête de la TDF figurent les 3 lignes des « *probas* » identiques à celles que nous avons vu au Chap. 1, 2.

Occ	T1	T2	T3	T4	T5	T6	T7	T8
18108	3627	781	2000	1742	939	2509	2702	3808
<i>p</i>	0,200	0,043	0,110	0,096	0,052	0,139	0,149	0,210
<i>q</i>	0,800	0,957	0,890	0,904	0,948	0,861	0,851	0,790

Telle est la première table de contingence qui dit *le status de la population recensée*.

1.3 La TDR

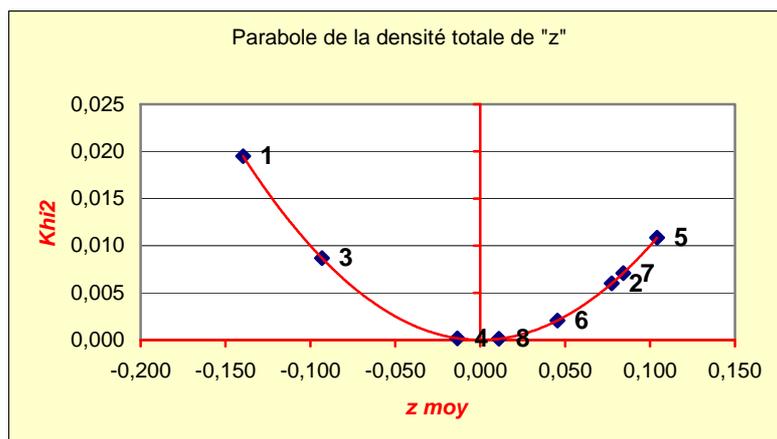
La TDR est à l'image de la TDF (Voir Chap. 1, 3), mais elle ne comprend que 10 colonnes. Colonne 1, le *Rang* (de 1 à 92, comme dans la TDF). Colonne 2, la somme ou la moyenne des densités. Colonnes 3 à 10, les densités des fréquences correspondant aux cases de la TDF.

Voyons les données correspondant aux lignes de la TDF présentées *supra in* Chap 5, 1.2 :

7,006	Tot	-12,845	7,133	-8,566	-1,229	9,575	4,187	7,746	1,006
0,076	Moy	-0,140	0,078	-0,093	-0,013	0,104	0,046	0,084	0,011
0,054	Khi2	0,019	0,006	0,009	0,000	0,011	0,002	0,007	0,000
1,000									
Ecart									
:	Moy	Max	Min		Borne inf	Borne sup			
	0,076	9,546	-5,357		2,500	-2,500			
					644	743	144,18 %		
Rang	Moy	T1	T2	T3	T4	T5	T6	T7	T8
1	-0,031	-0,925	-0,303	0,395	1,072	-0,708	-0,608	-0,426	1,254
2	0,085	4,001	0,885	1,230	-4,473	2,442	0,202	-0,188	-3,415
3	-0,014	0,892	0,002	1,262	-0,481	-0,125	-0,357	-1,402	0,097
4	0,209	-1,194	2,690	-1,760	9,546	-2,622	-3,323	-1,475	-0,189
5	-0,303	1,248	-1,324	-1,403	1,544	-2,209	-3,597	-2,621	5,939
6	-0,187	1,869	-2,353	-1,220	-0,765	-0,459	0,654	2,043	-1,262
7	-0,011	0,920	-0,886	0,253	1,148	0,703	-0,615	-2,066	0,458
8	0,018	-1,410	0,330	-0,061	0,799	-0,906	-1,349	3,334	-0,589
9	-0,014	1,308	-1,592	0,931	-5,357	1,579	5,918	2,427	-5,329
10	-0,024	0,425	0,603	-1,451	-0,557	-0,641	0,277	1,895	-0,742
83	0,075	-1,954	0,130	1,333	-1,496	1,157	-0,350	1,619	0,163
84	-0,027	-0,232	0,351	3,048	-1,334	-0,871	-0,896	-1,521	1,238
85	0,009	-1,553	-1,941	2,173	-1,184	1,768	1,303	-0,851	0,356
86	0,236	-0,271	3,425	-0,473	-2,175	2,531	-2,435	0,816	0,469
87	-0,049	-0,777	-0,685	0,639	0,957	0,210	-2,823	-0,883	2,971
88	0,123	1,475	4,044	0,518	-2,521	-0,897	0,990	-1,873	-0,752
89	0,114	1,139	1,508	-0,570	-0,581	0,564	1,413	-0,332	-2,225
90	-0,005	-0,804	-2,474	0,548	2,006	0,840	-0,060	1,270	-1,366
91	-0,004	-0,226	-1,036	2,259	0,702	-0,261	0,647	-1,589	-0,524
92	-0,288	2,750	-2,607	-0,422	0,016	-1,811	-0,684	-0,151	0,610

Le seuil de variation retenu est de $z = 2,500$ (correspondant à une probabilité de 99 %)

Les vertus de la TDR et des valeurs centrées réduites sont immédiatement exprimées par les images de synthèse, comme la parabole de la densité totale :



Il suffit d'observer pour voir. On voit d'emblée la position du tout et de la partie.

1.4 Les graphiques de base

La 4^{ème} page de la Macro ouvre un espace de graphiques qu'il ne convient pas de commenter ici pour des raisons évidentes. Tout graphique est destiné à illustrer des calculs algébriques et à faire voir les rapports et les relations entre les composantes mesurées.

1.5 La règle

La règle, à la 5^{ème} page de la Macro, ouvre un espace de lemmatisation et de discrimination remarquable. Elle est d'une utilité constante dans la description d'un corpus complexe ou multidimensionnel, puisqu'elle permet de mesurer la densité d'un vecteur composite (la *lemmatisation*) ou d'un vecteur isolé (la *discrimination*).

La mesure de densité, c'est la valeur centrée réduite « z » calculée au moyen de la formule classique déjà présentée au Chap. 1, 3.

La structure est une reprise de la première ligne de la TDR : en tête les « *probas* » indispensables aux calculs et dans les lignes suivantes on trouve les références d'identification des variables et, en gras, la ligne des données statistiques, somme et distribution des fréquences, dont la valeur centrée réduite est automatiquement affichée à la ligne suivante. La densité moyenne du vecteur est affichée dans la case sous le « total » des occurrences.

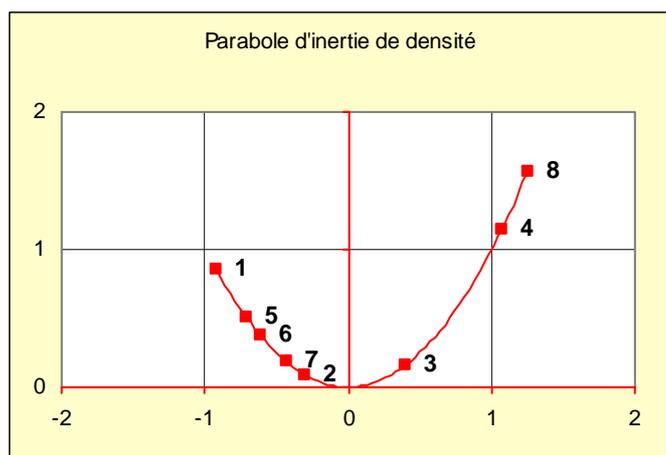
On peut encore dans la ligne en dessous soit dupliquer les valeurs de « z » pour tracer le nuage des points sous forme de droite (anamorphe) de Henry, soit afficher les carrés « z^2 » pour tracer la parabole polynomiale.

Voici la table des 8 contes de Perrault, avec en première ligne les valeurs correspondant aux valeurs de la première ligne de la TDF et de la TDR :

18108	3627	781	2000	1742	939	2509	2702	3808
<i>p</i>	0,200	0,043	0,110	0,096	0,052	0,139	0,149	0,210
<i>q</i>	0,800	0,957	0,890	0,904	0,948	0,861	0,851	0,790

Fréq	T1	T2	T3	T4	T5	T6	T7	T8
615	114	25	71	67	28	80	88	142
-0,250	-0,925	-0,303	0,395	1,072	-0,708	-0,608	-0,426	1,254
	0,856	0,092	0,156	1,149	0,501	0,370	0,182	1,572

La parabole, une fois tracée, est automatiquement mise à jour en fonction des valeurs calculées :



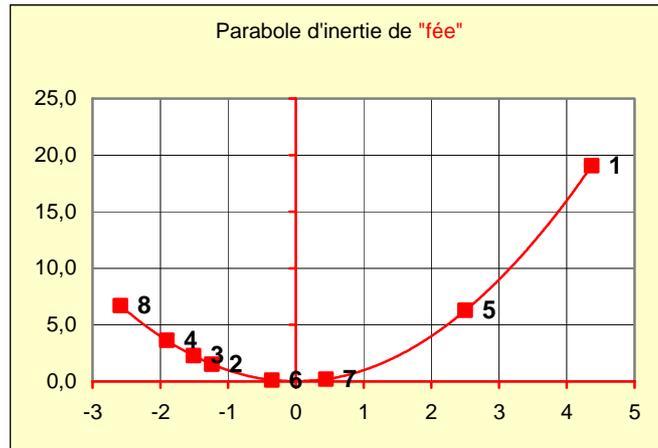
La *discrimination* et la *lemmatisation* se font à partir de la Base des données de la page *Lexique*.

1. La *discrimination* a pour objectif d'isoler un vecteur d'un ensemble composite, formé de plusieurs facteurs de même fréquence. On prend les valeurs brutes que l'on plaque dans la règle. Les calculs sont faits automatiquement et le graphique instantanément tracé.
2. La *lemmatisation* a pour objectif, au contraire, de rassembler un certain nombre de facteurs en un vecteur composite. Pour ce faire, le plus simple est d'abord de classer les données qualitatives par ordre alphabétique pour les voir se superposer, en règle générale ; et ensuite de les sélectionner et de les coller dans la page de la règle pour calculer la somme du vecteur recomposé et porter les données dans la règle.

Prenons par exemple le *lemme* des deux flexions de « *fée* » : *fée* et *fées*, regardons les valeurs calculées et voyons la parabole qui s'ensuit :

<i>fée</i>	21	7		1		4	3	6	
<i>fées</i>	13	10				1	1		1
TOTAL	34	17	0	1	0	5	4	6	1
		T1	T2	T3	T4	T5	T6	T7	T8
<i>z</i>	-0,273	4,366	-1,238	-1,507	-1,902	2,504	-0,353	0,446	-2,588
<i>z</i> ²		19,066	1,533	2,273	3,619	6,268	0,125	0,199	6,699

La parabole est instantanément tracée :



Force est de se rendre à l'évidence : le graphique traduit la qualité du *lemme* et renforce la pertinence de l'analyse statistique.

On voit :

- 1) que le lemme *fée* joue un rôle déterminant dans les contes 1 et 5
- 2) que la densité est autre chose que « le pourcentage » (ou *status*) de la population

À l'inverse, on peut isoler *par discrimination* le singulier *fée* du groupe des 21 fréquences qui l'englobent ou encore le pluriel *fées* du groupe des 13 occurrences qui l'agglutinent dans la TDF : il y a en effet 11 éléments à la fréquence 21 et 18 à la fréquence 13.

Toutes ces opérations sont effectuées d'un seul trait dans 5 premières pages de la Macro.

2. Table de corrélation, ACP et Métrique R

Une deuxième série d'opérations est disponible dans les pages 6 et 7 de la Macro.

2.1 La table de corrélation

Elle se fait d'un seul clic, automatiquement.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0,743	0,935	0,837	0,863	0,922	0,920	0,931
T2	0,743	1	0,776	0,711	0,724	0,734	0,701	0,711
T3	0,935	0,776	1	0,816	0,887	0,916	0,894	0,904
T4	0,837	0,711	0,816	1	0,727	0,779	0,828	0,879
T5	0,863	0,724	0,887	0,727	1	0,850	0,888	0,801
T6	0,922	0,734	0,916	0,779	0,850	1	0,883	0,869
T7	0,920	0,701	0,894	0,828	0,888	0,883	1	0,869
T8	0,931	0,711	0,904	0,879	0,801	0,869	0,869	1

C'est une matrice carrée, positive et symétrique, de trace $n (= 8)$, le nombre de variables du corpus.

On trouve en tête le vecteur de la moyenne marginale des corrélations, le \bar{r} qui est à la base de la métrique R exploitée par l'ACP dans la page suivante :

T1	T2	T3	T4	T5	T6	T7	T8
0,894	0,762	0,891	0,822	0,843	0,869	0,873	0,870

On sait déjà comment lire les valeurs de r en fonction du seuil $r = 0,866$, valeur limite de la liaison stochastique entre les variables (Voir Chap. 2,5 et Chap. 3, 5.1).

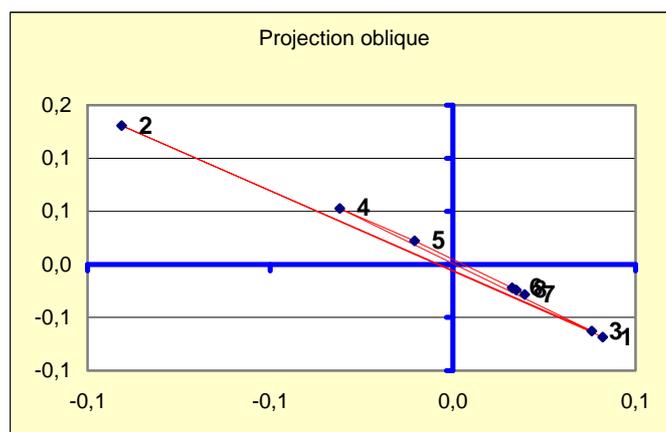
La valeur du *cosinus* r permet de formuler l'hypothèse de la *liaison stochastique* entre les variables appariées, sachant que la corrélation est intransitive. Inutile d'insister.

2.2 ACP et métrique R

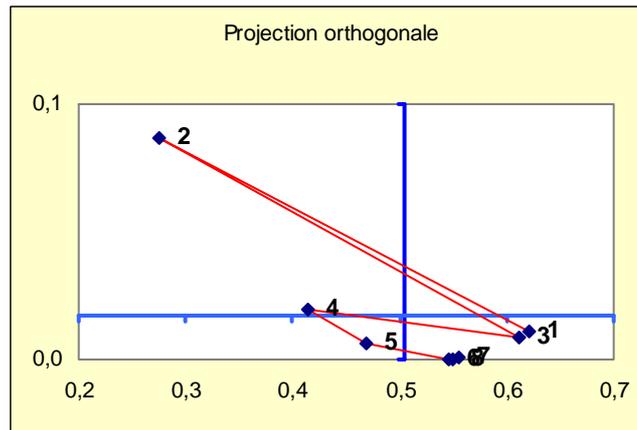
Pour ce qui est de la description de la structure faisant tous les calculs de la *métrique R*, voir au Chap. 2, *l'Ajustement*, la *Corrélation* et *l'ACP*. Comme on en connaît tous les paramètres, toutes les modalités de calcul et toutes les capacités graphiques, il est inutile d'insister.

Contentons-nous ici de visualiser les 3 graphiques fondamentaux concernant :

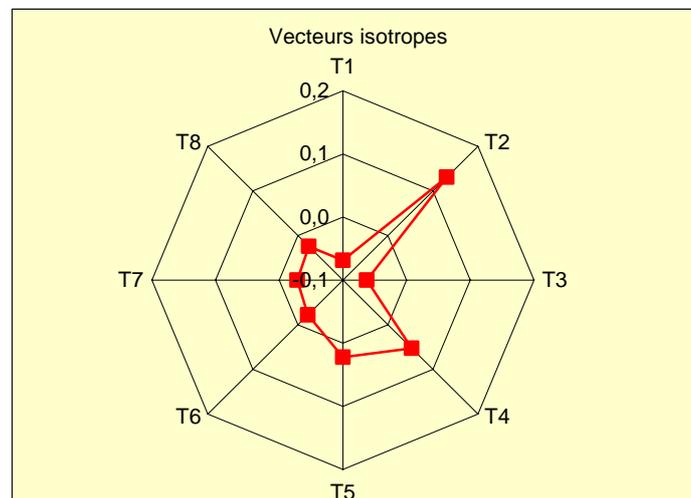
1. le nuage de points représentant la projection oblique (moment d'inertie vectoriel) :



2. le nuage de points représentant la projection orthogonale :



3. le nuage de points en radar qui met en évidence les éléments aberrants :



Point n'est besoin de grands discours pour voir comment les 8 contes sont liés entre eux, et remarquer que T2, *le Petit Chaperon rouge*, est poussé par une force centrifuge. Bien qu'étant le plus, il est porté par le vecteur le plus long.

L'ACP (Analyse en Composantes Principales) n'a d'autre finalité que de simuler la position des points descriptifs de la hiérarchie des variables. Les distances sont des dimensions sans mesure certes, mais elles portent en elles les « qualités inhérentes et essentielles » qui font qu'elles ont une puissance descriptive de la matrice des corrélations. Dans le cas présent, les « qualités inhérentes et essentielles » sont des qualités intrinsèques qui trouvent leur origine dans l'écriture et dans la composition des textes, exprimant une ligne de pensée continue et cohérente.

On peut imaginer, dans d'autres domaines, des qualités qui se réfèrent aux capacités curatives de médicaments ou de traitements expérimentés, à des contrôles de production, à des mesures écologiques, économiques ou financières...

Lorsque cette « vision » de la corrélation de masse est rigoureusement établie, on peut s'engager à coup sûr dans la comparaison (et l'appariement) des variables, on ne peut plus naviguer à vue. On sait de quoi on parle. On va tout droit « reconnaître » ce qu'on connaît déjà. On va progresser dans l'analyse qualitative des facteurs et des composantes vectorielles.

3. AFD et Estimation

C'est au terme du parcours descriptif mené à bien par la Macro qu'on touche à l'essentiel de l'analyse statistique. L'AFD (Analyse Factorielle Discriminante) s'exprime en trois dimensions (Voir Chap. 3) :

- 1) la régression linéaire simple
- 2) la régression multiple
- 3) la régression vectorielle ou factorielle

Comme les techniques de calcul de l'estimation et de l'ajustement qui entrent dans l'étude de la régression ont été exposées au chapitre 3, nous irons à l'essentiel et nous nous attarderons sur les principes de la régression multiple et de la régression factorielle qui font de l'AFD une analyse statistique performante et riche d'enseignements.

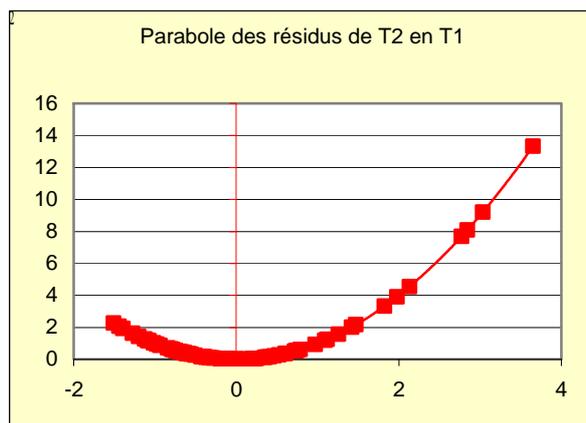
3.1 la régression linéaire simple

La régression linéaire simple consiste à appairer les variables 2 à 2 au gré des données statistiques produites par la TDF. Il s'agit de calculer Y' estimation de Y en X , et de transformer la décomposition de la variance pour en dégager les spectres de densité et les représentations spectrales (dont nous connaissons déjà les principes). Rien de nouveau malgré l'apparente complexité des données. Il y a une parfaite correspondance entre les lignes des 2 variables appariées, puisque la matrice est une matrice carrée ou rectangulaire pleine, même si le contenu de certaines cases y est nul. Elle exprime le *status* de la population recensée.

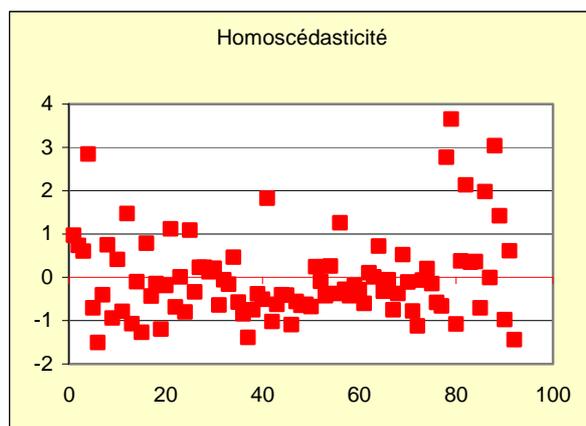
La TDF des 8 contes de Perrault comporte 92 lignes. Il s'ensuit que les sommes des carrés Vt^2 , Vr^2 et Vc^2 est égale à $n = 92$. (Voir Chap. 3, 6.2, comment les sommes des densités sont nulles et les sommes des carrés égales à n).

Dans l'estimation de T2 en T1, les données statistiques montrent que l'appariement des deux premiers contes est déficient. Avec un coefficient de corrélation $r = 0,743$, le taux de liaison n'est que de 33 %. Ce qui signifie que 30 des 92 lignes sont théoriquement liées et que 62 ne le sont pas (les éléments remarquables, aberrants ou déviants).

La parabole des résidus de T2 en T1 confirme le manque flagrant de liaison entre les textes des 2 premiers contes :



Distorsion encore confirmée par l'*homoscédasticité* des résidus (qui se caractérise par une dissémination ou un éparpillement des points sur un graphique difforme) :

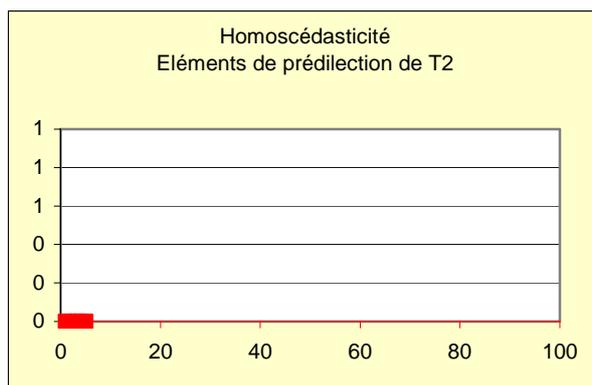


Oui, mais comment les identifier ? En utilisant les filtres adéquats proposés par Excel pour segmenter la plage des données et l'analyser par tranches. On isole ainsi les facteurs et les items en question que l'on identifie.

On peut sélectionner les lignes dont la valeur V_r du résidu est positive : $V_r \geq 0$. On dénombre alors 35 lignes.

On peut être encore plus exigeant en réduisant la valeur de dt : $dt \geq 0,5$. On ne dénombre plus que 14 lignes.

Limitation suprême, réduire les filtres pour ne sélectionner que les seuls éléments hautement significatifs (éléments remarquables) : $V_r \geq 1,96$ et $dt \geq 1$. On ne retient plus que 5 lignes sur les 92. Ce sont les éléments préférentiels qui caractérisent T2 par rapport à T1. Ce sont sans doute les éléments de prédilection de Perrault dans l'écriture de ce conte, qui ne cesse de se distinguer des autres.



Comme les lignes sont parfaitement identifiées, on n'a plus qu'à retourner au *Lexique* pour identifier les vocables et ensuite aux textes pour relever les séquences et les analyser comme il se doit. C'est le *summum* de l'analyse discriminante (AFD).

En effet, on les retrouve à la fréquence 5, 11, 14, 15 et 450.

À la fréquence 450, c'est l'article *le*, à valeur de déictique.

À la fréquence 11, 14 et 15 on trouve les éléments caractéristiques du *Petit chaperon rouge* : la *mère-grand*, le *loup* et le *chaperon rouge*.

À la fréquence 5, on trouve le *beurre*, la *galette* et le *pot*.

Sans autre commentaire, si nous voulons éviter de relever les séquences d'un texte bien connu.

Ajoutons, pour être complets, que l'une des structures de STABLEX permet de relever automatiquement les séquences en fonction des « mots » ou « vocables » ainsi identifiés, suivant l'étendue souhaitée ou exigée par les besoins de l'analyse.

Qui plus est, les séquences sont relevées et classées suivant le fil chronologique du texte. Cette technique est de la plus haute importance quand on veut suivre le fil du discours et le développement ou les mécanismes du raisonnement.

Il est difficile d'être plus précis et plus rigoureux. C'est toute la gamme des éléments qui est ainsi prise en compte : tout est mesuré, classé, répertorié, identifié, décrypté, analysé. On peut tout vérifier, on peut tout recommencer, les comptes sont (et seront) toujours les mêmes, rigoureusement identiques.

3.2 la régression linéaire multiple

La régression linéaire multiple consiste à estimer chaque variable du corpus par rapport à l'ensemble des variables représentées par la somme marginale des *Occurrences*. De fait, la régression multiple a, comme terrain de prédilection, la Base des données du *Lexique*. C'est là qu'elle va pouvoir comparer de manière efficace le contenu de chaque variable au contenu global du corpus.

Comment procéder ?

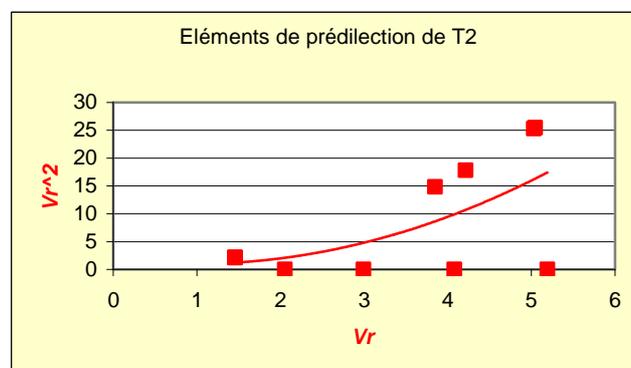
Sélectionner toute la matrice, et trier par ordre de fréquence décroissante les occurrences de la colonne qu'on veut analyser, et reporter le contenu dans une nouvelle feuille en lui accolant le contenu marginal correspondant. On ne retient du corpus que les seuls éléments qui se retrouvent dans la variable analysée. Il est évident que les éléments du corpus qui n'appartiennent pas à la variable sont sans intérêt et sans importance.

On reporte le tout dans une nouvelle feuille. On calcule les valeurs centrées réduites, puis on estime la variable en question en fonction des valeurs marginales et on ne retient que les variables des transformées.

Voyons, par exemple, les éléments de prédilection concernant le conte T2 du *Petit Chaperon rouge*, les éléments sélectionnés en fonction d'un $z > 0$, auquel correspond d'ailleurs un $Vr \geq 0$ avec un $dt \geq 1$:

Mot	Occ	T2	z	Vt	Vr	Vc	dt	Vr^2
Loup	14	14	17,624	2,684	5,046	-0,311	3,029	25,462
Mère-grand	15	14	16,971	2,684	5,030	-0,300	3,015	25,302
Petit Chaperon rouge	11	11	15,622	1,961	3,850	-0,345	2,356	14,824
beurre	5	5	10,532	0,513	1,459	-0,413	1,009	2,128
galette	5	5	10,532	0,513	1,459	-0,413	1,009	2,128
pot	5	5	10,532	0,513	1,459	-0,413	1,009	2,128
ma	39	10	6,556	1,720	2,993	-0,028	1,732	8,957
est	75	14	6,119	2,684	4,081	0,380	2,190	16,654
c'	46	8	4,366	1,237	2,053	0,052	1,159	4,216
le	450	31	2,690	6,785	5,193	4,627	1,081	26,965

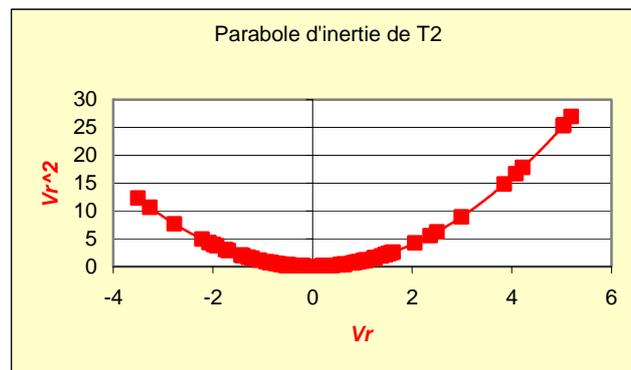
Lorsque les filtres sont maîtrisés et bien en place, les résultats sont sûrs. On retrouve à coup sûr les *items* choisis par l'auteur, le seul responsable de l'écriture et des choix de composition exprimés par les textes.



On peut adjoindre aux résultats des calculs tous les graphiques qui permettent d'en visualiser la distribution afin de guider correctement l'analyse discriminante.

On est en possession de tous les paramètres de la variable : le coefficient de corrélation pour déterminer le sens de la *liaison stochastique*, le taux d'éléments liés et le taux des éléments aberrants, les graphiques pour visualiser le nuage de points, et tous les facteurs qui font la trame du texte et conditionnent l'analyse.

On peut, juste à titre d'exemple, comparer la parabole représentative des 272 vocables qui constituent le lexique de T2 et la parabole des 92 lignes de l'estimation linéaire simple de T2 en T1 (voir *supra* § 3.1).



Bref, l'Analyse Factorielle Discriminante prend ainsi tout son sens et exprime avec force toute sa puissance analytique.

Les filtres opèrent de façon performante en fonction des seuils statistiques et des critères d'analyse retenus. Rien de plus simple. Rien de plus facile. Rien de plus performant. Rien de plus complet. Rien de plus exhaustif. Rien de plus sûr.

3.3 la régression factorielle ou vectorielle (par *lemmatisation* ou par *discrimination*)

La régression factorielle ou vectorielle permet de faire une AFD sur un vecteur composé d'éléments déterminé par lemmatisation ou par discrimination, dans le sens horizontal de la matrice (suivant la distribution des occurrences).

Prenons un exemple, la comparaison du vecteur des *fées* et du vecteur des *ogres* dans les 8 variables du corpus. On commence par « former » les 2 vecteurs (comme pour l'utilisation de la règle), puis on reporte les vecteurs-sommes dans la page d'Estimation de la Macro, en collant verticalement les valeurs « transposées ».

Étude de cas *typiques* et *atypiques* :

1. *comparaison absurde de lemmatisations disparates*

Pour comparer les emplois de « *fées* » et de « *ogres* » dans les 8 contes, on forme les 2 vecteurs-sommes à partir des données du Lexique :

a) le vecteur de « *fées* »

fée	21	7	0	1	0	4	3	6	0
fées	13	10	0	0	0	1	1	0	1
TOTAL	34	17	0	1	0	5	4	6	1

b) le vecteur de « *ogres* »

ogre	41	1	0	0	9	0	0	0	31
ogres	3	1	0	0	0	0	0	0	2
ogresse	7	6	0	0	0	0	0	0	1
ogresses	1	0	0	0	0	0	0	0	1
TOTAL	52	8	0	0	9	0	0	0	35

La comparaison porte ici sur les 8 variables qui totalisent 34 occurrences pour « *fées* » et 52 pour « *ogres* ».

Que disent les paramètres d'estimation ?

D'entrée de jeu le coefficient de corrélation $r = -0,104$ avec une valeur négative et proche de zéro, annonce une comparaison impossible : un contresens (de par la valeur négative) et une absurdité (de par la valeur pratiquement nulle). On stoppe net la comparaison.

2. comparaison de nombre par discrimination

Pour comparer les deux flexions de « *fée* », au singulier et au pluriel, on les isole :

a) le vecteur du singulier « *fée* »

fée	21	7	0	1	0	4	3	6	0
-----	----	---	---	---	---	---	---	---	---

b) le vecteur du pluriel « *fées* »

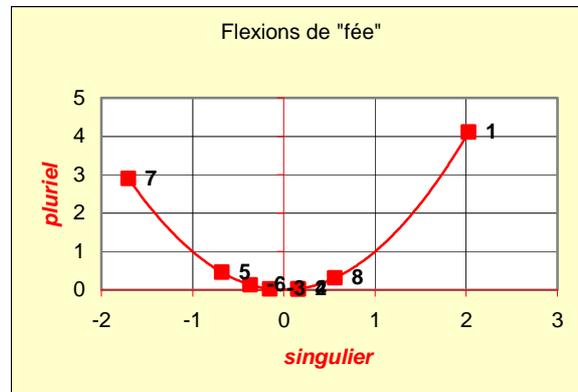
fées	13	10	0	0	0	1	1	0	1
------	----	----	---	---	---	---	---	---	---

Que disent les paramètres d'estimation ?

Avec une valeur $r = 0,634$, le coefficient de corrélation permet de formuler *une hypothèse de non liaison* entre l'emploi du singulier et du pluriel.

Que dit la parabole polynomiale ?

Elle traduit en image la réalité de faits. L'emploi du singulier et du pluriel est l'affaire du premier conte. Avec un résidu $Vr = 2,028$ il apparaît que l'emploi du pluriel est étroitement associé à l'emploi du singulier.



Mais la distance dt avec une valeur hautement significative, $dt = 1,462$ pour le conte 7 (T7), retient toute l'attention sur la valeur du résidu $Vr = -1,703$ significativement négatif de cette 7^{ème} variable.

Confirmant (ce qui n'est ici qu'une évidence) que l'emploi du pluriel « *fées* » est l'affaire de la variable Y , rangée du côté du singulier. En termes clairs : autant le singulier « *fée* » que le pluriel « *fées* » sont une caractéristique du premier conte, *La Belle au bois dormant* (que tout le monde connaît bien).

3. comparaison de genre et de nombre par discrimination et lemmatisation

Pour comparer les 2 vecteurs de « *ogre* », et différencier le genre et/ou le nombre, on les recompose :

a) le vecteur « *ogre* » au masculin, au singulier et au pluriel

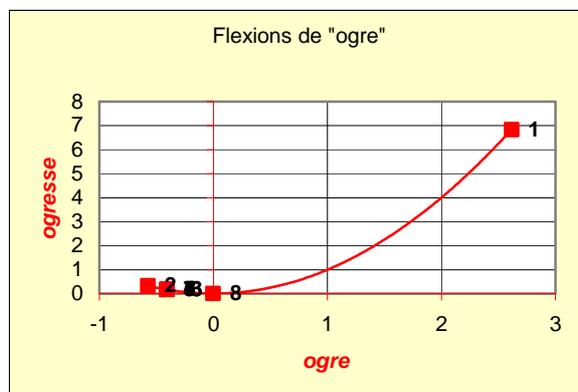
ogre	41	1	0	0	9	0	0	0	31
ogres	3	1	0	0	0	0	0	0	2
TOTAL	44	2	0	0	9	0	0	0	33

b) le vecteur « *ogresse* » au féminin, au singulier et au pluriel

ogresse	7	6	0	0	0	0	0	0	1
ogresses	1	0	0	0	0	0	0	0	1
TOTAL	8	6	0	0	0	0	0	0	2

Que disent les paramètres d'estimation ? Que dit la parabole ?

Le coefficient de corrélation est très faible : $r = 0,197$. L'hypothèse d'une absence de liaison annonce que l'emploi du substantif « *ogre* » répond aux exigences du discours, aux besoins de la thématique et à la dynamique de l'écriture.



La parabole, qui fait une place à part à la variable 1 de *La Belle au bois dormant*, montre que le nom « *ogre* » fait l'objet d'un emploi préférentiel dans ce premier conte.

Mais les valeurs spectrales traduisent la réalité des faits avec encore plus de précision, disant que l'emploi du féminin est l'affaire de T1 et l'emploi du masculin, celle de T8 :

1. la valeur spectrale $V_r = 2,615$ du résidu et la valeur totale de V_t (de Y) = 2,500 sont hautement significatives de T1. Et donc, l'emploi du féminin « *ogresse* » est un emploi préférentiel de T1, *La Belle au bois dormant*.
2. la valeur de la distance $dt = 1,767$ est hautement significative de T8, le dernier conte. Mais ce sont les valeurs spectrales du résidu $V_r = -0,002$ et de V_c (de X) = 2,548 qui font de l'emploi du masculin « *ogre* » une caractéristique de T8, *Le Petit Poucet*.

Voilà comment les spectres mesurent la réalité des faits et comment les images spectrales les mettent en vue.

4. comparaison de type normal

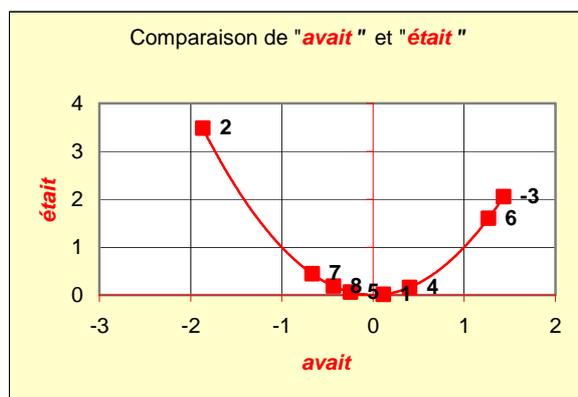
Mais l'étude des cas typiques ou atypiques ne peut nullement occulter la portée des mesures spectrales dans l'étude des *appariements normaux*, même lorsqu'il s'agit des vecteurs linéaires transversaux.

Comparons l'emploi des verbes « *avait* » et « *était* » :

avait	43	2	15	14	6	21	21	28	43
était	37	7	20	9	8	24	18	24	37

Les paramètres sont *normaux* : $r = 0,946$, hypothèse d'une relation linéaire forte. Le taux de liaison est de 5 éléments sur 8. Il y a donc 3 éléments aberrants.

C'est ce qui ressort de la parabole d'inertie :



On voit que les 5 éléments liés et que les 3 éléments aberrants sont clairement désignés et parfaitement positionnés sur les 2 branches de la parabole. Disant que l'emploi de « *était* » l'emporte sur celui de « *avait* » dans T3 et T6, mais qu'il est perdant dans T2.

En effet, les valeurs des résidus V_r et des distances dt font ressortir les deux moments forts de cette comparaison entre « *avait* » et « *était* », faisant de « *était* » un temps fort de T3, *la Barbe-bleue*, et de « *avait* » un temps fort de T4, *le Chat botté*.

Les résidus de T6 montrent que « *était* » occupe une place de choix dans *Cendrillon*.

T	X	Y	V_t	V_r	V_c	dt
1	2	7	-1,184	0,401	-1,389	0,586
2	15	20	0,169	1,433	-0,311	1,030
3	14	9	-0,976	-1,866	-0,394	1,158
4	6	8	-1,080	-0,247	-1,057	0,077
5	21	24	0,586	1,265	0,187	0,804
6	21	18	-0,039	-0,666	0,187	0,489
7	28	24	0,586	-0,433	0,767	0,465
8	43	37	1,939	0,113	2,011	0,384

Ce sont absolument tous les paramètres qui servent de guide à l'analyse discriminante.

Imaginons que pour des raisons techniques on veuille comparer les temps des verbes, les personnes grammaticales, des séries de pronoms, des vecteurs synonymiques, etc., on tient la recette : d'abord on isole les vecteurs, ensuite on les mesure, puis on considère les paramètres et on observe les graphiques, et enfin on se lance dans l'analyse, c'est-à-dire dans

l'explication des raisons de la partie et du tout, de la partie dans le tout et du tout face à la partie.

Cet échantillon montre parfaitement combien l'*AFD* place l'estimation linéaire au cœur de l'analyse statistique, l'analyse discriminante par excellence. Quel plaisir n'a-t-on pas dans ces conditions à analyser un corpus tout entier !

3.4 la régression linéaire simple

On effectuera la régression linéaire simple à partir des données de la TDF. Mais, dans ce cas, on ne perdra jamais de vue les densités fournies par la TDR.

3.5 Les paramètres utiles pour l'*AFD* (la régression linéaire multiple)

On précisera dans la 8^{ème} chapitre la formation des cônes inversés sur lesquels se projette la spirale représentative de la distribution des éléments en fonction des densités qui caractérisent tous et chacun des éléments constitutifs.

Pour l'heure, disons que la feuille de la MACRO contenant les valeurs de X , Y , Y' et z auxquelles sont accolées les valeurs correspondantes de V_t , V_r , V_c , dt et V_r^2 , doit être paramétrée :

- 1) on sélectionne d'abord les valeurs de $V_r \geq 0$ avec un dt *décroissant* pour avoir les valeurs correspondant aux facteurs de prédilection ayant un $dt \geq 1$ pour les éléments de la dominante thématique et stylistique, ayant un $0,500 \leq dt < 1$ pour les éléments de la sous-dominante thématique et stylistique et ayant un $0 < dt < 0,500$ pour les éléments de développement ou d'embellissement thématique et stylistique. Cette partie, comme on le verra au chapitre 8 aux §5 et 9, correspond au cône isotrope supérieur.
- 2) on sélectionne ensuite les valeurs de $V_r \leq 0$ avec un dt *croissant* auquel correspond le cône isotrope inversé, pendant du cône supérieur.

On se laisse alors guider par les valeurs ainsi paramétrées pour procéder aux différentes analyses exigées par la qualité du lexique, du texte et du discours insérés dans le texte.

Telle est l'*AFD*, l'*Analyse Factorielle Discriminante*, qui représente comme un aboutissement de l'analyse statistique appliquée à l'analyse lexicale, textuelle et discursive.

On trouvera une ébauche d'application de l'*ACP* et de l'*AFD* à l'analyse lexicale, textuelle et discursive du conte *Le Petit Chaperon rouge*, au chapitre 9, du présent ouvrage.

4. Conclusion

La statistique à la portée de tous prend ici tout son sens, un sens que l'on ne pouvait même pas imaginer il y a quelques temps encore. C'est grâce à la puissance du calcul informatique que le calcul statistique est devenu une réalité à la portée de tous.

C'est par la pratique de la méthode statistique que l'on découvrira les subtilités de l'*AFD*, l'analyse factorielle discriminante, que l'on entrera dans les arcanes de l'analyse, au sens propre du terme, que l'on percera le mystère des rapports quantitatifs et qualitatifs entre la partie et le tout et le tout dans la partie.

Une métaphore suffirait à l'expliquer. Qu'est-ce qu'une montre ? Une montre, ce n'est pas un ensemble de pièces pêle-mêle, mais l'ensemble des pièces qui ont pour vocation de

fonctionner ensemble dans un tout. Pourquoi faire ? Ce n'est certainement pas pour « donner l'heure », comme on le dit ou comme on le croit. La montre ne donne rien, même pas l'heure. La montre donne à son utilisateur le « centre trigonométrique » de l'espace et du temps qu'il occupe et qui le préoccupent. C'est, à l'image du GPS, un guide spatio-temporel.

De la statistique pratique à la pratique de la statistique, il n'y a qu'un pas qu'on franchit allègrement. La statistique est à la portée de tous.

La méthode révèle les secrets de la statistique et la statistique les secrets de la méthode.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

6

Mode d'emploi de la statistique analytique

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce sixième chapitre, notre intention est de proposer un *Mode d'emploi de la statistique analytique*, avec toujours la ferme intention de mettre la statistique à la portée de tous.

D'un point de vue pratique, nous allons mettre l'accent sur les mesures essentielles ou fondamentales indispensables pour mener à bien une analyse statistique discriminante.

D'un point de vue didactique, nous allons aborder l'étude d'un corpus de données médicales que nous empruntons encore une fois à G. SAPORTA dans *Probabilités, Analyse des Données et Statistique*, publié aux Editions TECHNIP, Paris, 1990, p. 413-414, (suivant des données communiquées par M. J. P. Nakache).

D'un point de vue technique, nous allons porter notre attention sur l'analyse exploratoire d'une base de données complexes et multidimensionnelle, en tournant les pages de la Macro.

Pour ce faire, nous allons d'abord présenter cette base de données médicales, ensuite nous irons à l'essentiel des tables de contingence (TDF et TDR), enfin nous nous attarderons sur l'analyse factorielle discriminante (AFD) pour voir comment les spectres nous disent l'essentiel des corrélations phénoménales que les composantes entretiennent entre elles. Certes, personne n'attend ici de réponse médicale, ça va de soi. En revanche, tout le monde attend de voir comment la méthode statistique perce les secrets d'un corpus, montre les rapports inhérents aux composantes et fait voir les caractéristiques essentielles à travers les phénomènes décrits. La statistique, c'est faire connaître pour faire reconnaître, de façon cohérente, objective, descriptive et inductive. Elle fait passer de l'induction des phénomènes à l'essence des phénomènes.

1. Base des données

La base des données est tirée de G. Saporta, *op. cit.*, p. 413-414.

Constitution de la base de données. 101 victimes d'infarctus du myocarde ont subi un examen clinique au moment de leur hospitalisation. Ils ont été soumis à 7 paramètres d'analyse qui constituent les 7 variables du corpus :

1. *frcar* – fréquence cardiaque
2. *incard* – index cardiaque
3. *insys* – index systolique
4. *prdia* – pression diastolique
5. *papul* – pression artérielle pulmonaire
6. *pvent* – pression ventriculaire
7. *repul* – résistance pulmonaire

Le tableau est complété par une 8^{ème} colonne indiquant la *survie* ou le *décès* des patients, que nous plaçons en tête. Sur les 101 victimes d'infarctus, 51 sont mortes et 50 ont survécu.

Tout l'intérêt de ce corpus est, du point de vue statistique, de définir à l'intérieur des 7 paramètres d'examen clinique quels sont les critères qu'il convient de retenir du point de vue médical pour expliquer soit la survie soit le décès.

1.1 Le tableau des données

Voilà la retranscription fidèle du tableau complet retenu par G. Saporta. Ce sont les 101 cas soumis aux 7 critères d'analyse, avec la reproduction des mesures médicales enregistrées :

<i>N°</i>	<i>S/D</i>	<i>frcar</i>	<i>incard</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
1	survie	90	1,71	19	16	19,5	16	912
2	décès	90	1,68	18,7	24	31	14	1476
3	décès	120	1,4	11,7	23	29	8	1657
4	survie	82	1,79	21,8	14	17,5	10	782
5	décès	80	1,58	19,7	21	28	18,5	1418
6	décès	80	1,13	14,1	18	23,5	9	1664
7	survie	94	2,04	21,7	23	27	10	1059
8	survie	80	1,19	14,9	16	21	16,5	1412
9	survie	78	2,16	27,7	15	20,5	11,5	759
10	survie	100	2,28	22,8	16	23	4	807
11	survie	90	2,79	31	16	25	8	717
12	survie	86	2,7	31,4	15	23	9,5	681
13	survie	80	2,61	32,6	8	15	1	460
14	survie	61	2,84	47,3	11	17	12	479
15	survie	99	3,12	31,8	15	20	11	513
16	survie	92	2,47	26,8	12	19	11	615
17	survie	96	1,88	19,6	12	19	3	809
18	survie	86	1,7	19,8	10	14	10,5	659
19	survie	125	3,37	26,9	18	28	6	665
20	survie	80	2,01	25	15	20	6	796

21	survie	82	3,15	38,4	13	20	6	508
22	décès	110	1,66	15,1	23	31	6,5	1494
23	décès	80	1,5	18,7	13	17	12	907
24	décès	118	1,03	8,7	19	27	10	2097
25	décès	95	1,89	19,9	25	27	20	1143
26	décès	80	1,45	18,1	19	23	15	1269
27	décès	85	1,3	15,1	13	18	10	1108
28	décès	105	1,84	17,5	18	22	10	957
29	survie	122	2,79	22,9	25	36	10	1032
30	survie	81	1,77	21,9	18	27	11	1220
31	survie	118	2,31	19,6	22	27	10	935
32	décès	87	1,2	13,8	34	41	20	2733
33	décès	65	1,19	18,3	15	18	13	1210
34	survie	84	2,15	25,6	27	37	10	1377
35	décès	103	0,91	8,8	30	33,5	10	2945
36	survie	75	2,54	33,9	24	31	16	976
37	survie	90	2,08	23,1	20	28	6	1077
38	survie	90	1,93	21,4	11	18	10	746
39	décès	90	0,95	10,6	20	24	6	2021
40	survie	65	2,38	36,6	16	22	12	739
41	décès	95	0,99	10,4	20	27,5	8	2222
42	décès	95	0,85	8,9	19	22	15,5	2071
43	survie	86	2,05	23,8	21	28	10	1093
44	survie	82	2,02	24,6	16	22	14	871
45	décès	70	1,44	20,6	19	26,5	11	1472
46	survie	92	3,06	33,3	10	15	6	392
47	décès	94	1,31	13,9	26	40	15	2443
48	décès	79	1,29	16,3	24	31	10	1922
49	survie	67	1,47	21,9	15	18	16	980
50	décès	75	1,21	16,1	19	24	4	1587
51	survie	80	2,41	30,9	19	24	7	797
52	survie	61	3,28	54	12	16	7	390
53	décès	110	1,24	11,3	22	27,5	11	1774
54	décès	116	1,85	15,9	33	42	13	1816
55	survie	75	2	26,7	16	22	5	880
56	décès	92	1,97	21,4	18	27	3	1096
57	survie	110	0,96	8,8	15	19	16	1583
58	survie	95	2,56	26,9	8	13	3	406
59	survie	75	2,32	30,9	8	10	6	345
60	survie	80	2,65	33,1	13	19	9	574
61	décès	102	1,6	15,7	24	31	16	1550
62	survie	86	1,67	19,4	18	23	8,5	1102
63	décès	60	0,82	13,7	22	32	13	3122
64	survie	100	1,76	17,6	23	33	2	1500
65	survie	80	3,28	41	12	17	2	415
66	survie	108	2,96	27,4	24	35	6,5	946
67	décès	92	1,37	14,8	25	46	11	2686
68	décès	100	1,38	13,8	20	31	11	1797
69	survie	80	2,85	35,6	25	32	7	898
70	décès	87	2,51	28,8	16	24	20	765
71	survie	100	2,31	23,1	8	12	1	416
72	décès	120	1,18	9,9	25	36	8	2441

73	décès	115	1,83	15,9	25	30	8	1311
74	survie	101	2,55	25,2	23	30,5	9	957
75	survie	92	2,17	23,5	19	24	3	885
76	décès	87	1,42	16,1	20	26	10	1465
77	survie	80	1,59	19,9	13	20,5	4	1031
78	décès	88	1,47	16,7	23	32,5	10	1769
79	décès	104	1,23	11,8	27	33	11	2146
80	survie	90	1,45	16,1	17	24	8,5	1324
81	décès	67	0,85	12,7	26	33	11	3106
82	survie	87	2,37	27,2	15	22	10	743
83	survie	108	2,4	22,2	26	31	4	1033
84	décès	120	1,91	15,9	18	27	15	1131
85	décès	108	1,5	13,9	28	43	16	1813
86	survie	86	2,36	27,4	24	34	8	1153
87	décès	112	1,56	13,9	24	29	4	1487
88	décès	80	1,34	17	16	25	16	1493
89	décès	95	1,65	17,4	20	33	7	1600
90	décès	90	2,04	22,7	28	41	10	1608
91	survie	90	3,03	33,6	17	23,5	7	620
92	décès	94	1,21	12,9	17	22	3	1455
93	décès	51	1,34	26,3	11	17	6	1015
94	décès	110	1,17	10,6	29	35	10,5	2393
95	décès	96	1,74	18,1	24	29	6	1333
96	décès	132	1,31	9,9	23	28	12	1710
97	décès	135	0,95	7	15	20	7	1684
98	décès	105	1,92	18,3	18	24	3	1000
99	décès	99	0,83	8,4	23	27	8	2602
100	décès	116	0,6	5,2	33	38	10	5067
101	décès	112	1,54	13,8	25	31	8	1610

Il est important de conserver toutes les mesures, puisque ce sont précisément elles qui vont nous permettre de mener à bien l'analyse statistique qui devrait permettre au praticien de conforter le pronostic médical.

Rappelons que la statistique ne se prononce pas sur l'essence des phénomènes analysés, mais qu'elle se cantonne dans le rôle d'une servante fidèle pour un diagnostic sûr, grâce à sa puissance analytique et descriptive.

1.2 Le traitement statistique

Le tableau des données constitue la matrice statistique des 101 cas traités (classés de 1 à 101) soumis aux 7 critères d'analyse médicale. Tous les paramètres d'analyse statistique sont ainsi réunis pour un traitement correct du corpus.

Comment procéder ?

On forme la base des données, telle que présentée ci-après :

- colonne 1 : identification du cas
- colonne 2 : somme des mesures médicales pour chaque cas

– colonnes 3 à 9 : la distribution des 7 critères d’analyse.

S/D	Total	frcar	incar	insys	prdia	papul	pvent	repul
survie	1074,21	90	1,71	19,0	16	19,5	16,0	912
décès	1655,38	90	1,68	18,7	24	31,0	14,0	1476
décès	1850,1	120	1,40	11,7	23	29,0	8,0	1657
survie	929,09	82	1,79	21,8	14	17,5	10,0	782
décès	1586,78	80	1,58	19,7	21	28,0	18,5	1418
décès	1809,73	80	1,13	14,1	18	23,5	9,0	1664
survie	1236,74	94	2,04	21,7	23	27,0	10,0	1059
survie	1561,59	80	1,19	14,9	16	21,0	16,5	1412
...

On colle cette base de données (c’est la matrice de données) dans la page 1 de la Macro, en conservant la ligne des titres. La cellule **S/D** va à la cellule **A2**, la cellule **Total** à la cellule **B2**, et ainsi de suite. De sorte que les 9 colonnes sont correctement remplies par les mesures effectuées sur les malades, lesquels sont parfaitement identifiés.

On lance l’analyse statistique que la Macro effectue d’un seul trait dans les 5 premières pages et au cas par cas dans les 3 dernières. Suivons-en le déroulement.

2. Les tables de contingence (TDF et TDR)

La Macro produit les tables de contingence que nous connaissons bien maintenant, la table de distribution des fréquences (TDF) et la table des écarts centrés réduits (TDR).

2.1 La TDF

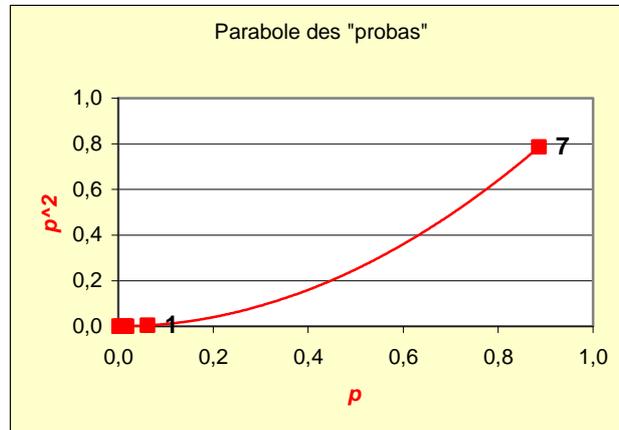
La TDF reprend dans le cas présent les données brutes de la matrice d’entrée, puisqu’on ne compte qu’un seul individu par ligne.

En tête de table figurent les valeurs de probabilité « p » et « q » :

	101		150857,32	9308	186,42	2102,4	1945	2626	959,5	133730
	7		p	0,062	0,001	0,014	0,013	0,017	0,006	0,886
			q	0,938	0,999	0,986	0,987	0,983	0,994	0,114
Rang	Occ	Nbre	Fréq	frcar	incar	insys	prdia	papul	pvent	repul
1	1074,21	1	1074,21	90	1,71	19	16	19,5	16	912
2	1655,38	1	1655,38	90	1,68	18,7	24	31	14	1476
100	5269,8	1	5269,8	116	0,6	5,2	33	38	10	5067
101	1801,34	1	1801,34	112	1,54	13,8	25	31	8	1610

Commentaires.

1. La TDF reproduit le *status* de la population. Il faut le considérer tel quel.
2. Les valeurs de probabilité donnent la parabole d'inertie de « p » :



Le nuage de points est sans équivoque. On remarque qu'il y a 6 paramètres agglutinés et un paramètre qui se détache, celui de *repul*, l'indice 7 de la *résistance pulmonaire*.

S'il est vrai que la quantité d'information fournie par la probabilité est l'inverse du logarithme de « p », alors on peut voir que les quantités varient d'une variable à l'autre :

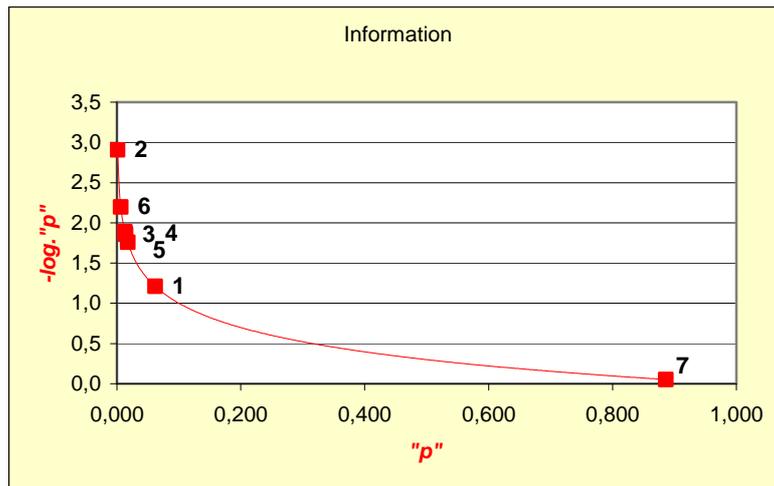
	1	2	3	4	5	6	7
<i>variable</i>	frcar	incar	insys	prdia	papul	pvent	repul
<i>-log.p</i>	1,210	2,908	1,856	1,890	1,759	2,197	0,052

Les critères 2 et 6, *index cardiaque* et *pression ventriculaire*, sont les deux paramètres qui fournissent le plus d'information. On ne les perdra pas de vue.

Viennent ensuite les critères 3, 4 et 5, *index systolique*, *pression diastolique* et *pression artérielle pulmonaire*, avec des indices élevés.

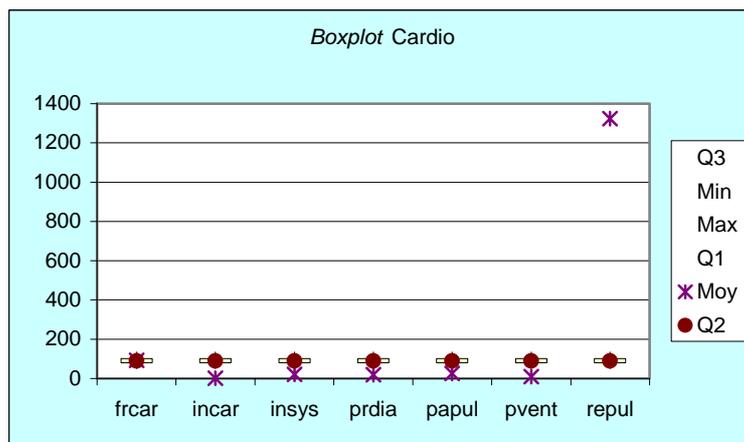
Le critère 1 de la *fréquence cardiaque* affiche un indice moindre.

Enfin, on note que l'indice du 7 de la *résistance pulmonaire* (*repul*) tend à « déclasser » ce critère d'analyse, qui occupe une place à part.



Même s'il s'agit de simples constatations, le praticien peut d'entrée de jeu les prendre en considération. Toutes ces informations servent de guide pour une meilleure approche de la réalité des faits.

3. la *boxplot* de Tukey est sans intérêt :



Que montre-t-elle ? On voit que l'étoile de la moyenne de 7 (*repul*) est projetée en haut du graphique, ce qui singularise à l'extrême cette variable de la *résistance pulmonaire*, comme s'il s'agissait d'un paramètre totalement étranger à l'affaire.

L'étoile de la moyenne de 1 (*frcar*) est dans la boîte, comme si la *fréquence cardiaque* était un paramètre banal ou banalisé.

Puis, entre les deux extrêmes, l'étoile de la moyenne des paramètres 2 à 6, *index cardiaque*, *index systolique*, *pression diastolique*, *pression artérielle pulmonaire* et *pression ventriculaire*, est en dessous de la boîte.

Bref, autant d'indications qui sont clairement données par la parabole d'inertie des « probas », où le 1 de *frcar* se rapproche du groupe des 5 et le 7 de *repul* s'en éloigne considérablement.

Au gré des informations, on voit que les paramètres ne sont pas tous sur un pied d'égalité.

La TDF est le passage obligé pour entrer dans le domaine des densités et des corrélations.

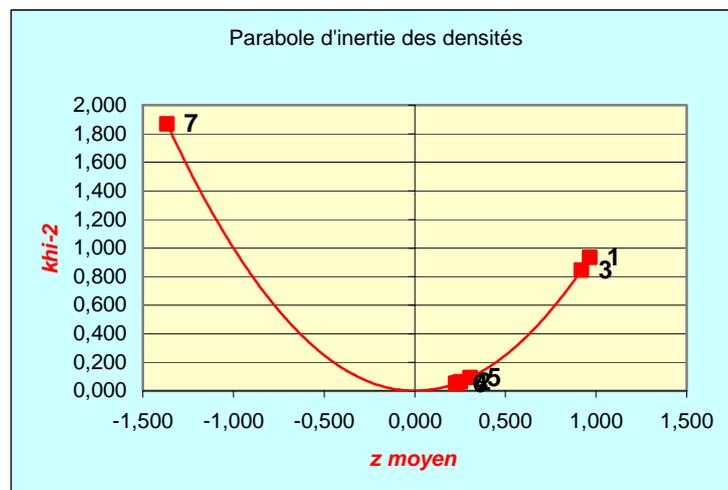
2.2 La TDR

La TDR est la transformation de la TDF en table des valeurs centrées réduites (les écarts réduits « z »).

155,255	Total	97,608	25,378	92,892	23,531	30,986	22,907	-138,048
1,537	Moy	0,966	0,251	0,920	0,233	0,307	0,227	-1,367
3,911	Khi2	0,934	0,063	0,846	0,054	0,094	0,051	1,868
0,790								
Ecart :	Moy	Max	Min		Borne inf	Borne sup		
	1,537	17,173	-12,990		-2,000	2,000		
					126	253	46,91%	
Rang	Moy	frcar	incar	insys	prdia	papul	pvent	repul
1	0,686	3,008	0,332	1,049	0,582	0,187	3,519	-3,871
2	0,045	-1,240	-0,256	-0,916	0,579	0,411	1,073	0,663
100	-2,734	-11,974	-2,318	-8,019	-4,267	-5,660	-4,075	17,173
101	-0,341	0,084	-0,460	-2,272	0,371	-0,064	-1,025	0,978

Considérations classiques :

1. La parabole d'inertie des densités (moyenne des « z » et khi-2) :

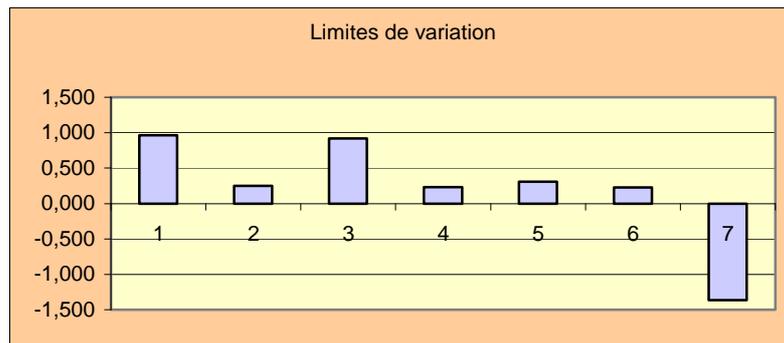


La parabole dit tout. Il n'est nullement besoin de commentaires pour voir :

- qu'il y a une variable remarquable ou « aberrante », la 7 qui est rejetée à gauche,
- qu'il y a 4 variables centrées et proches, les 2, 4, 5 et 6,
- et qu'il y en a 2 de proches, la 3 et la 1, qui sont d'une importance non négligeable.

Sachant que lesdites « variables » représentent les tests cliniques que subissent les patients au moment de l'hospitalisation, ils méritent d'être considérés avec attention.

2. Histogramme de la distribution et limites de normalité :



Le seul intérêt de cet histogramme est de montrer les limites de variation et le degré de normalité de la distribution. 6 critères poussent vers le haut et le 7^{ème} vers le bas.

3. La normalité de la distribution (khi-2 de Fisher)

Avec un khi-2 de 3,911 à 7 ddl et une probabilité de 79 %, les limites de la distribution sont repoussées bien au-delà de ce que montre le graphique. C'est la variable n° 7 (*repul*) qui pose problème dans le groupe. Néanmoins, avec une valeur de $-1,367$, le 7 reste dans la norme avec une probabilité de 99%. En effet, avec une probabilité de 98 % à 7 ddl, les limites de normalité de distribution sont repoussées à $\pm 1,564$.

Or la plus forte valeur du groupe (celle du 7) est toujours dans l'intervalle de définition limité à $\pm 1,500$ comme il apparaît dans le graphique ci-dessus.

La distribution est normale puisqu'elle est contrôlée à 100%.

L'analyse statistique peut alors se poursuivre normalement, les densités sont de toute évidence significatives.

4. Lectures de la TDR

La lecture de la TDR doit être une lecture croisée (horizontale et verticale). Les valeurs sont immédiatement décryptées en fonction des seuils de densité. Les valeurs normales sont dans l'intervalle $-1,96 \leq z \leq +1,96$, arrondies à ± 2 . Et les valeurs remarquables ou « aberrantes » sont à l'extérieur de cet intervalle.

Pour repérer les vecteurs, on accolera sur la droite du tableau dans les colonnes vides les paramètres d'identification, comme le numéro d'ordre et la fin du parcours médical (*survie* ou *décès*).

5. Quelques observations

En faisant varier dans la Macro les bornes des seuils de probabilité de la TDR, on voit les paramètres cliniques découper des zones opposées dans les cas de *survie* et dans les cas de *décès* :

- a) Si la variation réduite est réduite à zéro, les 5 premiers paramètres (*frcar*, *incar*, *insys*, *prdia*, *papul*) vont de pair et s'opposent au dernier paramètre (*repul*). Lorsque les premiers sont en rouge (positifs), le dernier est en bleu (négatif) ; en règle générale, il s'agit des *cas de survie*. Dans le cas contraire, les premiers en bleu et le dernier en rouge, il s'agit des *cas de décès*.
- b) En mettant les limites de z sont à ± 2 , seuls les paramètres 1 et 3 (*frcar* et *insys*) vont de pair et s'opposent au 7^{ème} (*repul*). Dans les *cas de survie*, les colonnes sont de couleur rouge, rouge et bleu. Dans les *cas de décès*, les couleurs des colonnes sont inversées : bleu, bleu et rouge.

Si on ne perd pas de vue la parabole d'inertie des « *probas* », on voit que la 7^{ème} variable est toujours en opposition.

Comme il s'agit de valeurs hautement significatives, il y a tout lieu de penser que les paramètres d'observation clinique ne sont pas tous de la même importance. Que certains méritent sans doute d'être privilégiés. Il faut faire varier les seuils de probabilité pour analyser plus finement encore les contrastes :

- a) Si l'intervalle est réduit à zéro, les 5 premiers paramètres vont de pair et s'opposent au 7^{ème}
- b) Si l'intervalle est à ± 2 , le 1 (*frcar*) et le 3 (*insys*) vont de pair et s'opposent au 7^{ème} (*repul*)
- c) Si l'intervalle est à $\pm 2,5$, même constat, bien que le 3^{ème} paramètre (*insys*) tende à s'effacer
- d) Si l'intervalle est à ± 3 , même constat
- e) Si l'intervalle est à ± 4 , les paramètres 1 (*frcar*), 3 (*insys*) et 7 (*repul*) opposent clairement *les cas de survie* et *les cas de décès*. Les autres paramètres sont pratiquement neutralisés.

Voilà quelques observations qui montrent combien la TDR dit le tout et la partie, conformément au principe du calcul algébrique qui réduit la partie à l'intégralité.

En revanche, l'ACP et l'AFD sont plus que jamais nécessaires pour pouvoir comparer et analyser les variables sur toute l'étendue factorielle.

Pour ce faire, il faut reprendre la Macro et activer les trois dernières pages (de la *Corrélation*, de la *Métrique R* et de la *Régression*).

3. Les analyses croisées (ACP et AFD)

L'analyse en composantes principales se fait en 2 temps :

- 1) Dans un premier temps, on essaie de voir et de comprendre quelle est la corrélation entre les variables, c'est l'ACP
- 2) Dans un deuxième temps, on estime les vecteurs appariés en fonction des besoins de l'AFD (analyse factorielle)

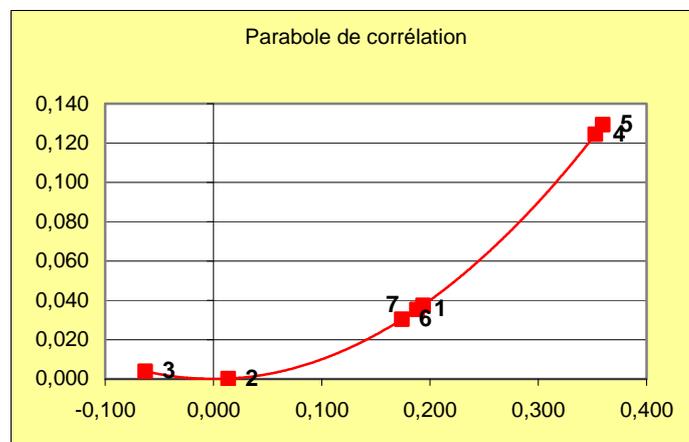
3.1 La table de corrélation et les liaisons stochastiques

	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repu</i>
	0,188	0,014	-0,063	0,353	0,360	0,174	0,194
	0,035	0,000	0,004	0,124	0,129	0,030	0,037
<i>frcar</i>	1	-0,112	-0,503	0,399	0,370	-0,085	0,247
<i>incar</i>	-0,112	1	0,887	-0,361	-0,269	-0,282	-0,767
<i>insys</i>	-0,503	0,887	1	-0,483	-0,405	-0,201	-0,735
<i>prdia</i>	0,399	-0,361	-0,483	1	0,928	0,285	0,702
<i>papul</i>	0,370	-0,269	-0,405	0,928	1	0,244	0,650
<i>pvent</i>	-0,085	-0,282	-0,201	0,285	0,244	1	0,258
<i>repu</i>	0,247	-0,767	-0,735	0,702	0,650	0,258	1
	1	2	3	4	5	6	7

Le coefficient r moyen de la corrélation est très faible, compte tenu du niveau de signification ($r \geq 0,866$). (Voir Chap. 2 et 3)

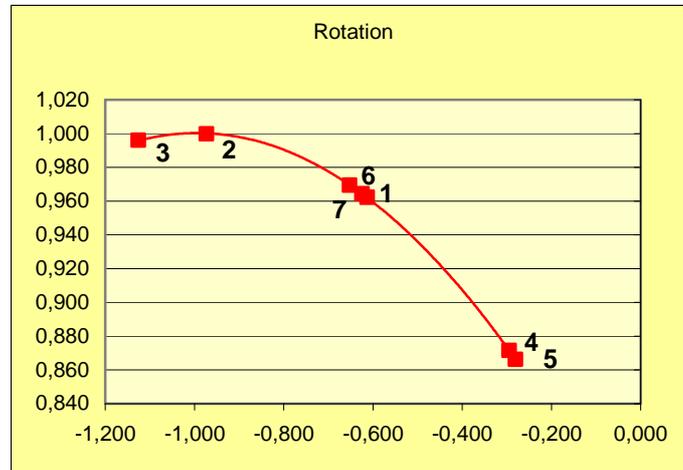
Néanmoins, on voit qu'il y a 2 couples de corrélations significatives : le couple des paramètres 2 et 3, *incar* et *insys*, avec un $r = 0,887$ et le couple des paramètres 4 et 5, *prdia* et *papul*, avec un $r = 0,928$.

Rapprochement confirmé par la parabole d'inertie de la corrélation :



Mais la parabole met aussi en évidence le voisinage des paramètres 1, 6 et 7 (*frcar*, *pvent* et *repul*).

Image confirmée par les rotations de la métrique R :



Comme toutes les images convergent, il est inutile de les multiplier.

3.2 ACP et AFD (lectures croisées)

Si l'ACP fournit une indication sur la position des variables, qui est en soi une indication de la liaison stochastique, l'AFD va permettre de poursuivre l'analyse factorielle par une lecture transversale des variables appariées (spectre de transformation et images de synthèse).

L'estimation prend tout son sens dans la valeur du coefficient de corrélation r qui définit le sens et le degré de la *liaison stochastique* des variables comparées. Cette valeur donne un sens à l'hypothèse de travail qu'elle oriente.

Est-ce que les rapprochements des *index cardiaque* et *index systolique*, 2 et 3, (*incar* et *insys*), des *pressions diastolique* et *artérielles pulmonaire*, 4 et 5, (*prdia* et *papul*) ou encore de la *fréquence cardiaque*, de la *pression ventriculaire* et de la *résistance pulmonaire*, 1, 6 et 7, (*frcar*, *pvent* et *repul*) sont dus au hasard ?

Ou bien est-ce que l'analyse statistique doit retenir l'attention du praticien sur des phénomènes dont l'importance est plus grande qu'il n'y paraît à première vue ?

L'estimation et l'examen exhaustif des données sont rendus possibles en fonction de l'appariement des variables parce qu'il permet de comparer les mesures cliniques effectuées au moment de l'hospitalisation et définir la portée et l'importance des paramètres cliniques.

En fonction des seuils d'observation qui permettent d'apprécier l'espérance de vie du patient, il doit être possible de se faire une opinion sur les chances de vie et de formuler un diagnostic probable. Si tel est le cas, la statistique joue pleinement son rôle de description et d'aide à l'interprétation des données.

3.2.1 Considérations sur les chances de « décès » ou de « survie »

Les observations portent toutes sur les densités de la TDR ou du spectre de décomposition des variables.

Les filtres de la TDR permettent de séparer les densités de « décès » des densités de « survie » et de ranger les valeurs par ordre de fréquence croissant ou décroissant.

C'est une espèce de *lemmatisation* qui permet de sélectionner directement dans la TDR le tableau des éléments de « décès » et le tableau des éléments de « survie » (ce sont des sous totaux).

3.2.1.1 densités et paramètres en cas de « décès » :

a) Le corpus compte 51 cas de « décès »

La TDR détache en bleu les densités négatives et en rouge les densités positives en fonction des paramètres cliniques :

Décès		1	2	3	4	5	6	7
N°	Moy	frcar	incar	insys	prdia	papul	pvent	repul
100	-2,734	-11,974	-2,318	-8,019	-4,267	-5,660	-4,075	17,173
63	-1,734	-10,284	-1,601	-4,746	-3,115	-3,321	-1,708	12,636
81	-1,702	-9,754	-1,583	-4,886	-2,483	-3,174	-2,141	12,107
35	-1,532	-6,699	-1,505	-5,311	-1,643	-2,870	-2,229	9,536
99	-1,516	-5,672	-1,402	-4,893	-2,138	-3,079	-2,297	8,870
41	-1,218	-4,434	-1,140	-3,987	-1,949	-2,192	-1,845	7,022
72	-1,168	-3,474	-1,154	-4,466	-1,561	-1,484	-2,153	6,120
67	-1,154	-6,623	-1,159	-4,022	-1,997	-0,580	-1,711	8,015
39	-1,151	-3,928	-1,059	-3,601	-1,523	-2,267	-2,110	6,432
24	-1,051	-1,977	-1,066	-4,124	-1,931	-2,034	-1,187	4,964
94	-1,025	-4,064	-1,135	-4,272	-0,764	-1,513	-1,475	6,052
42	-0,998	-3,759	-1,150	-4,010	-1,835	-2,728	0,347	6,150
32	-0,936	-7,201	-1,273	-4,260	-0,618	-1,413	0,317	7,900
47	-0,915	-5,546	-1,078	-3,790	-1,373	-0,870	-0,429	6,680
97	-0,897	1,892	-0,895	-3,758	-1,865	-2,217	-1,422	1,986
79	-0,833	-3,442	-0,975	-3,660	-0,567	-1,207	-1,001	5,021
50	-0,720	-3,152	-0,633	-1,634	-0,695	-1,113	-2,113	4,300
6	-0,704	-3,093	-0,740	-2,230	-1,111	-1,438	-0,742	4,426
48	-0,693	-4,512	-0,801	-2,380	-0,556	-0,883	-0,896	5,177
92	-0,682	-0,522	-0,550	-2,016	-0,817	-1,134	-2,263	2,527
53	-0,608	-1,010	-0,758	-3,080	-0,648	-1,135	-0,412	2,790
68	-0,590	-2,040	-0,679	-2,633	-1,088	-0,579	-0,441	3,331
3	-0,489	0,565	-0,586	-2,793	-0,176	-0,570	-1,102	1,242
78	-0,488	-2,995	-0,600	-2,003	-0,407	-0,222	-0,669	3,482
96	-0,453	1,307	-0,688	-3,275	-0,345	-0,936	-0,054	0,817
89	-0,390	-1,427	-0,366	-1,483	-0,605	0,385	-1,279	2,048

87	-0,361	0,902	-0,352	-1,960	0,531	-0,018	-2,040	0,409
101	-0,341	0,084	-0,460	-2,272	0,371	-0,064	-1,025	0,978
76	-0,292	-1,371	-0,416	-1,387	-0,211	-0,435	-0,106	1,879
22	-0,255	0,635	-0,290	-1,733	0,286	0,323	-1,286	0,278
45	-0,238	-3,096	-0,398	-0,421	-0,417	-0,325	0,217	2,775
88	-0,233	-2,222	-0,489	-1,255	-1,147	-0,696	1,709	2,469
85	-0,139	-1,556	-0,633	-2,712	0,377	1,322	0,875	1,354
27	-0,106	0,923	-0,197	-0,561	-0,782	-0,814	0,728	-0,039
33	-0,103	-2,010	-0,363	-0,089	-0,553	-1,114	1,537	1,868
54	-0,082	-0,896	-0,421	-2,362	1,321	1,106	0,011	0,671
95	-0,041	0,317	-0,090	-0,640	1,041	0,542	-1,163	-0,296
93	-0,026	-2,299	-0,045	2,689	-0,934	-0,599	-0,439	1,444
61	-0,023	-0,536	-0,376	-1,749	0,332	0,129	1,487	0,550
2	0,045	-1,240	-0,256	-0,916	0,579	0,411	1,073	0,663
90	0,060	-2,073	-0,125	-0,484	0,996	1,736	-0,433	0,803
26	0,098	-0,876	-0,235	-0,399	0,146	-0,368	1,977	0,442
73	0,102	2,359	-0,023	-1,120	1,273	0,743	-0,513	-2,003
56	0,126	1,674	0,332	0,925	0,440	1,094	-1,776	-1,811
5	0,156	-1,868	-0,272	-0,517	0,121	0,073	2,655	0,900
98	0,195	3,985	0,394	0,497	0,755	0,811	-1,634	-3,443
23	0,370	1,958	0,179	1,074	-0,144	-0,298	2,069	-2,246
84	0,450	4,334	0,209	-0,613	0,211	0,812	2,260	-4,060
28	0,476	4,349	0,374	0,440	0,900	0,524	1,049	-4,301
25	0,771	1,461	0,191	0,313	1,902	0,800	3,974	-3,247
70	1,559	3,897	1,246	4,348	1,108	1,887	5,734	-7,309

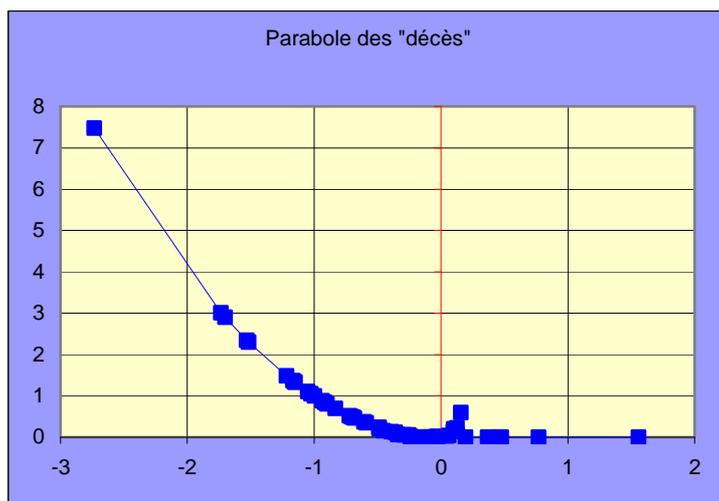
On observe en tête de tableau que les densités des paramètres 1 à 6 sont en bleu et que celle du 7^{ème} est en rouge. Puis en fin de tableau, un glissement vers l'inversion des couleurs des densités qui font exception à la règle.

On observera néanmoins les valeurs de la colonne de droite concernant la 7^{ème} variable de la *résistance pulmonaire (repul)* qui ne compte que 8 cas d'exception en bleu, alors que toutes les autres densités sont en rouge.

Ce tableau offre la possibilité d'apprécier les paramètres cliniques à leur juste valeur (chose que le praticien ne manquera pas de faire).

b) La parabole d'inertie de l'ensemble des cas (AFD)

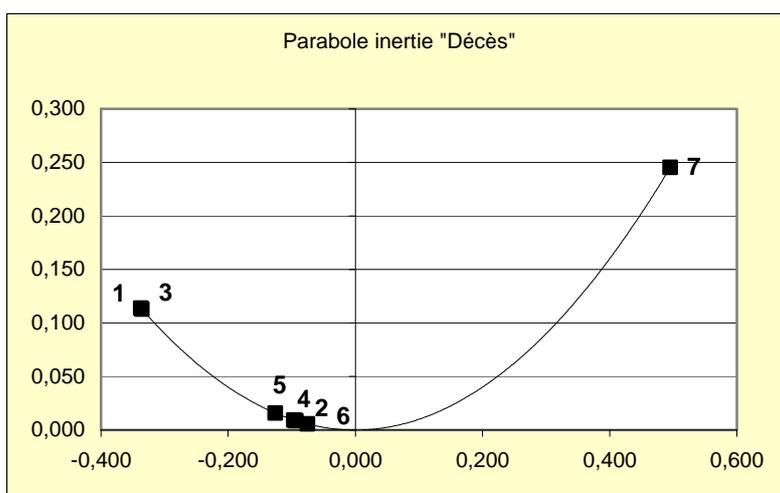
La parabole d'inertie reflète la synthèse de la distribution des décès : la « normalité » des valeurs négatives (39 cas de décès sur les 51 recensés) sont sur la branche négative gauche et les 12 cas d'exception à la règle (12 sur 51) sont sur la branche positive de droite.



c) La synthèse des « probas » des densités de « décès »

Décès		1	2	3	4	5	6	7
Paramètres cliniques		frcar	incar	insys	prdia	papul	pvent	repul
Total	100534,93	4891	71,13	764,8	1120	1484	543	91661
Moyenne	1971,27	95,90	1,39	15,00	21,96	29,10	10,65	1797,27
Densité	-28,999	-17,198	-4,767	-17,119	-4,926	-6,415	-3,826	25,252
Moyenne	-0,580	-0,337	-0,093	-0,336	-0,097	-0,126	-0,075	0,495
Moy^2	0,336	0,114	0,009	0,113	0,009	0,016	0,006	0,245

d) La parabole d'inertie (ACP)



On voit combien l'ACP et l'AFD vont de pair : elles offrent une lecture croisée des densités.

3.2.1.2 densités et paramètres en cas de « survie »

Le corpus compte 50 cas de « survie » sur les 101 recensés et observés.

a) *Le corpus compte 50 cas de « survie ».*

La TDR détache en bleu les densités négatives et en rouge les densités positives en fonction des paramètres cliniques :

Survie		1	2	3	4	5	6	7
N°	Moy	frcar	incar	insys	prdia	papul	pvent	repul
57	-0,553	0,184	-0,820	-3,184	-1,609	-2,102	1,458	2,201
64	-0,346	-0,355	-0,217	-1,203	0,297	0,710	-2,662	1,006
80	-0,238	-0,149	-0,281	-1,006	-0,483	-0,354	-0,301	0,909
8	-0,222	-1,720	-0,533	-1,481	-0,927	-1,196	2,091	2,210
77	-0,054	0,949	0,120	0,896	-0,540	0,030	-1,266	-0,567
62	0,165	0,978	0,092	0,447	0,443	0,235	0,176	-1,216
30	0,181	-0,468	0,049	0,610	0,047	0,610	0,751	-0,332
17	0,307	4,927	0,637	1,711	-0,110	0,563	-1,262	-4,316
34	0,352	-1,306	0,158	0,825	1,536	1,895	0,019	-0,664
37	0,366	1,543	0,435	1,385	0,988	1,366	-0,686	-2,473
49	0,488	-0,257	0,074	1,606	0,150	-0,339	3,339	-1,157
43	0,506	0,938	0,391	1,484	1,173	1,290	0,694	-2,426
55	0,521	1,511	0,650	3,299	0,764	0,985	-0,601	-2,965
75	0,529	3,503	0,768	2,341	1,500	1,357	-1,426	-4,342
7	0,568	2,091	0,414	1,083	1,778	1,190	0,763	-3,346
86	0,592	0,415	0,554	2,054	1,648	2,253	-0,169	-2,607
83	0,604	3,835	0,719	1,244	2,578	2,106	-1,365	-4,891
20	0,646	2,943	0,781	3,288	0,816	0,888	-0,002	-4,189
10	0,653	5,302	0,980	2,516	0,973	1,476	-0,887	-5,792
4	0,678	3,364	0,599	2,477	0,588	0,333	1,688	-4,303
1	0,686	3,008	0,332	1,049	0,582	0,187	3,519	-3,871
38	0,719	4,794	0,779	2,528	-0,172	0,603	1,799	-5,294
18	0,760	5,371	0,714	2,603	-0,103	0,015	2,402	-5,687
31	0,818	5,929	0,768	0,962	1,943	1,649	1,041	-6,569
44	0,872	2,374	0,660	2,715	0,745	0,962	2,913	-4,269
74	0,921	3,698	0,950	2,315	2,144	2,372	0,630	-5,664
29	0,969	5,269	1,002	1,319	2,224	3,076	0,727	-6,836
51	1,037	2,783	1,124	4,822	1,893	1,797	0,362	-5,522
9	1,044	2,972	0,971	4,223	0,943	1,162	2,367	-5,329
66	1,051	4,541	1,292	2,862	2,398	3,379	-0,302	-6,815
71	1,075	11,444	1,938	5,490	0,280	0,713	-1,367	-10,973
82	1,081	4,288	1,181	4,127	0,975	1,579	1,769	-6,349
69	1,243	1,686	1,312	5,331	2,985	3,069	0,049	-5,733
13	1,264	7,305	2,174	8,450	0,099	1,427	-1,445	-9,166
11	1,287	4,890	1,613	5,319	1,346	2,438	0,987	-7,584
36	1,297	0,430	0,927	4,450	2,361	2,434	3,190	-4,716
16	1,336	6,552	1,539	4,878	0,625	1,494	2,728	-8,464
40	1,369	1,377	1,216	6,895	1,331	1,652	2,661	-5,548
12	1,372	4,800	1,614	5,732	1,235	2,160	1,772	-7,710

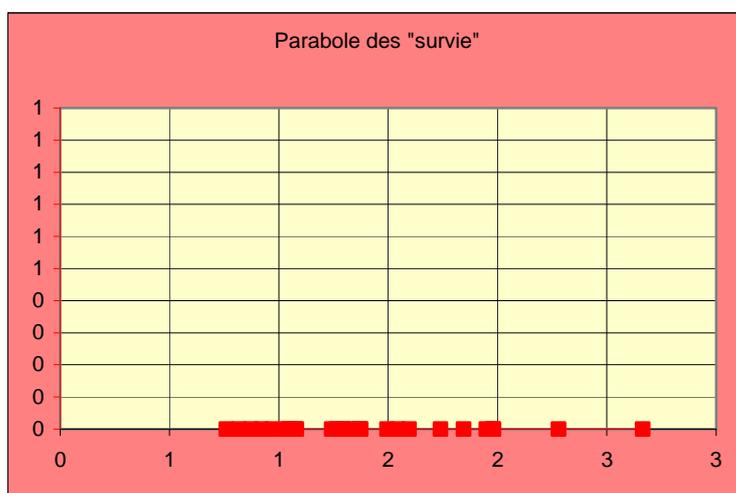
58	1,374	10,729	2,266	6,946	0,320	1,087	-0,281	-11,447
19	1,496	10,017	2,209	4,258	2,027	3,318	0,193	-11,552
60	1,544	5,368	1,840	7,231	1,173	1,776	2,025	-8,604
91	1,595	6,047	2,069	6,821	2,127	2,626	0,870	-9,393
59	1,740	8,667	2,255	9,469	0,750	0,593	1,707	-11,261
21	1,845	6,520	2,552	9,571	1,491	2,459	0,843	-10,520
46	1,950	10,262	2,884	9,306	1,091	1,759	1,336	-12,990
65	1,980	7,799	3,070	11,807	1,725	2,265	-0,857	-11,950
15	1,983	8,880	2,448	7,176	2,043	2,306	3,150	-12,124
14	2,279	3,662	2,337	13,089	1,015	1,837	4,005	-9,995
52	2,664	4,900	3,186	16,992	1,900	2,146	1,913	-12,388

Pour lire le tableau, il faut se laisser guider par les couleurs de densités positives (en rouge) et négatives (en bleu).

Point n'est besoin de grands commentaires pour voir qu'il se dégage de cet ensemble de valeurs une « norme » qui montre qu'en cas de « survie », les 6 premiers paramètres cliniques sont en rouge (positifs) et le 7^{ème} en bleu (négatif).

En outre, le tableau met en lumière les cas particuliers et les exceptions.

b) La parabole de l'ensemble des cas (AFD)



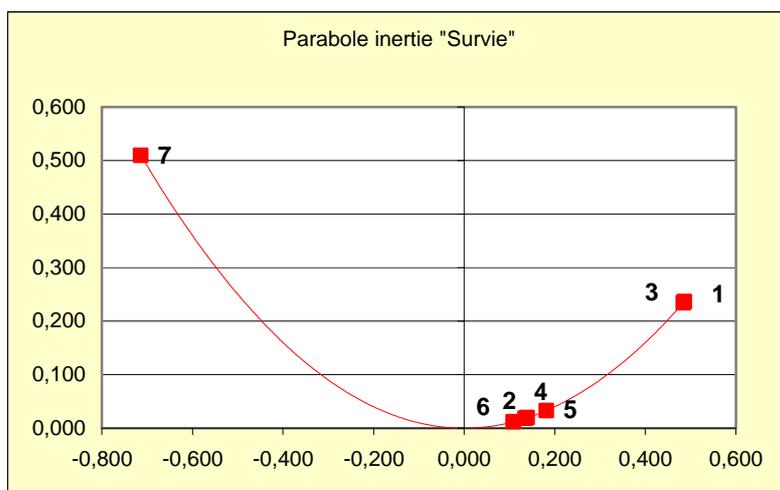
La parabole d'inertie montre la configuration exacte de la distribution qui s'étale sur la branche positive de droite, à l'exception des 5 cas particuliers qui sont sur la branche négative de gauche.

Tout est dénombré, tout est recensé, tout est mesuré, tout est connu et reconnu. Pas un seul élément, pas un seul facteur ne manque à l'appel.

c) La synthèse des « probas » des densités de « survie »

Survie		1	2	3	4	5	6	7
Paramètres cliniques		frcar	incar	insys	prdia	papul	pvent	repul
Total	50322,39	4417	115,29	1337,6	825	1142	416,5	42069
Moyenne	1006,45	88,34	2,31	26,75	16,50	22,84	8,33	841,38
Densité	40,988	24,309	6,738	24,196	6,962	9,068	5,407	-35,693
Moyenne	0,820	0,486	0,135	0,484	0,139	0,181	0,108	-0,714
Moy^2	0,672	0,236	0,018	0,234	0,019	0,033	0,012	0,510

d) La parabole d'inertie (ACP)



Il suffit de rapprocher les 2 paraboles de « décès » et de « survie » pour observer l'ampleur du phénomène et projeter les mécanismes d'analyse qui doivent aborder les mesures dans leur vraie dimension et scruter le problème médical dans son essence même.

3.2.1.3 Images de synthèse

Pour avoir une vision synthétique de la relation entre cas de « décès » et cas de « survie », il suffit de mettre en parallèle les valeurs correspondantes et d'en faire une image de synthèse.

1) les valeurs et les mesures des densités

Rappel : les probabilités « p » et « q » sont les probabilités du corpus tout entier. Les quantités d'information sont les mêmes et la parabole d'inertie est la même.

a) les mesures sont en cas de « décès » le double des mesures en cas de « survie » avec des différences notables entre les paramètres cliniques comme il ressort du tableau comparatif :

D/S	Total	frcar	incar	insys	prdia	papul	pvent	repul
décès	100534,93	4891	71,13	764,8	1120	1484	543	91661
survie	50322,39	4417	115,29	1337,6	825	1142	416,5	42069
paramètres		1	2	3	4	5	6	7

b) les densités sont globalement négatives en cas de « décès » et positives en cas de « survie », mais avec des valeurs algébriques de signes opposés (couleurs opposées) :

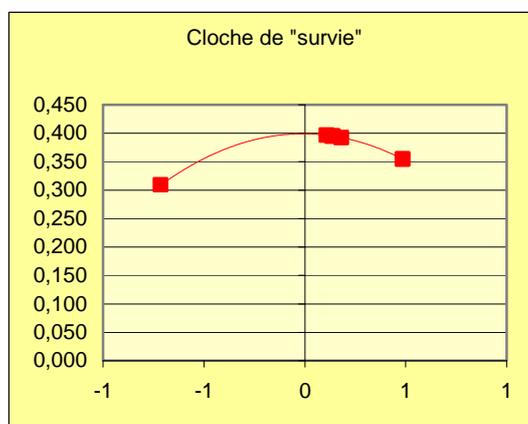
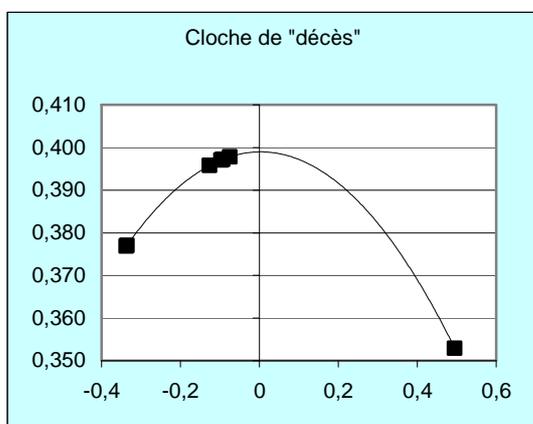
D/S	Moy	frcar	incar	insys	prdia	papul	pvent	repul
décès	-0,580	-0,337	-0,093	-0,336	-0,097	-0,126	-0,075	0,495
survie	0,820	0,486	0,135	0,484	0,139	0,181	0,108	-0,714
paramètres		1	2	3	4	5	6	7

2) les images de synthèse

Inutile de multiplier les images et les graphiques, il convient de choisir ce qui « parle » le mieux à l'esprit.

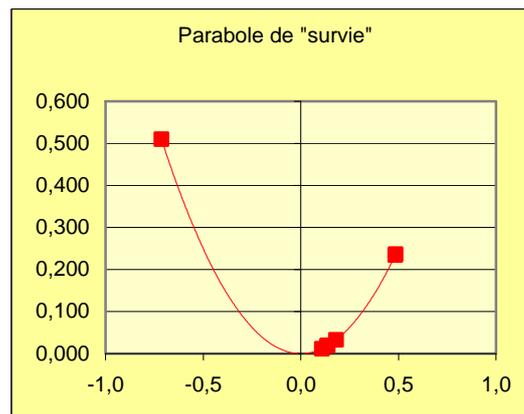
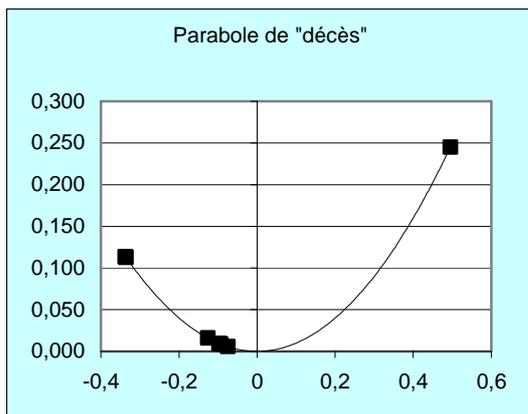
a) les paraboles d'inertie

À valeurs inversées, courbes inversées :



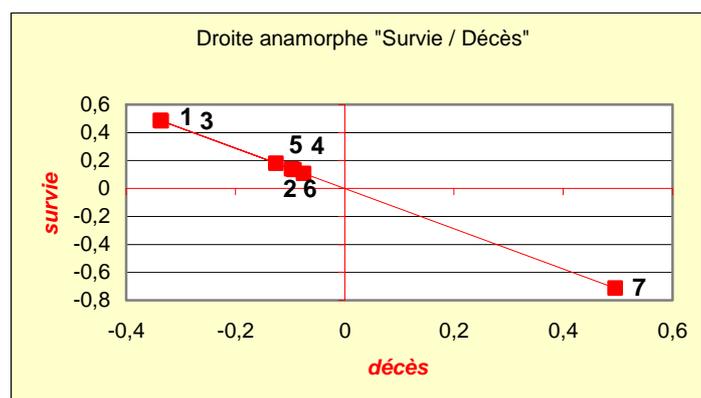
b) les paraboles

Les paraboles simulent mieux que les cloches la réalité des distributions :



c) la droite « anamorphe » de Henry

L'anamorphose de la droite de Henry est un moyen commode de représenter la distribution de part et d'autre du centre de gravité :



La position des points du nuage montre que les 6 premiers paramètres cliniques sont contrebalancés par le 7^{ème} :

- le paramètre 7 (*repul*) contrebalance les 6 autres
- les paramètres 1 (*frcar*) et 3 (*insys*) surclassent les 4 autres paramètres « alignés »
- les 4 autres paramètres tendent vers le centre de gravité de la moyenne 0 : le 2 (*incar*), le 4 (*prdia*), le 5 (*papul*) et le 6 (*pvent*)

C'est ainsi que l'ACP et l'AFD jouent un rôle complémentaire évident dans la description d'un corpus. Les valeurs marginales qui servent de référence pour les calculs poussent à la lecture croisée de la matrice des données.

La régression linéaire simple ou multiple, considérée verticalement ou transversalement, provoque nécessairement une lecture croisée des matrices de données. (Voir Chap. 5, 3)

4. Les analyses spectrales des résidus (ASR)

La régression linéaire est de la plus haute importance dans la mesure où le spectre de transformation et de décomposition de la variance révèle les caractéristiques des résidus de la liaison stochastique.

La TDR donne la densité des composantes factorielles et vectorielles.

Le spectre, lui, donne la densité de décomposition et de transformation de l'ajustement et de l'estimation des 2 variables appariées, aussi bien dans la régression linéaire simple que dans la régression linéaire multiple, ou encore dans la régression linéaire transversale (provenant de la lemmatisation). (Voir Chap. 5, 3)

Tout l'intérêt est de décrypter la densité des résidus qui portent en eux toute la puissance de la liaison linéaire (densité et intensité) avec une parfaite identification des éléments factoriels ou vectoriels. C'est là qu'on mesure à quel point la partie s'intègre dans le tout et à quel point le tout tient la partie.

Dans le cas présent, nous allons directement à la lecture croisée de la matrice des données en prenant comme variables les 2 sous-vecteurs de « décès » et de « survie » précédemment constitués.

4.1 Matrice des données des 2 variables de « décès » et de « survie »

Les filtres de lemmatisation ont permis de refaire une nouvelle matrice de données avec les 2 variables de « décès » et de « survie » pour axes verticaux, les 2 sous-ensembles de 51 cas pour l'un et de 50 cas pour l'autre. Quant aux facteurs linéaires, ce sont les 7 paramètres cliniques.

paramètres		Total	Décès	Survie
1	<i>frcar</i>	9308	4891	4417
2	<i>incar</i>	186,42	71,13	115,29
3	<i>insys</i>	2102,4	764,8	1337,6
4	<i>prdia</i>	1945	1120	825
5	<i>papul</i>	2626	1484	1142
6	<i>pvent</i>	959,5	543	416,5
7	<i>repul</i>	133730	91661	42069
Total		150857,32	100534,93	50322,39

D'où la nouvelle matrice de « données rassemblées » que la Macro pourrait traiter en tant que telle. Néanmoins, comme tous les paramètres d'analyse sont déjà connus, nous allons directement à l'analyse spectrale des résidus (l'ASR)

4.2 les estimations Y' de Y (survie) en X (décès)

Le module d'estimation de la Macro donne tous les renseignements nécessaires pour avoir une parfaite connaissance de la corrélation entre les 2 sous-parties du corpus.

4.2.1 Spectre de transformation

Le tableau du spectre de transformation concernant l'estimation Y' de Y (*survie*) en X (*décès*) les densités de V_t , V_r et V_c , suivies des distances dt (du khi-2) et des carrés de V_t , V_r et V_c :

paramètres		V_t	V_r	V_c	dt	V^2	V_r^2	V_c^2
1	<i>frcar</i>	-0,194	2,229	-0,300	1,584	0,038	4,967	0,090
2	<i>incar</i>	-0,495	-0,903	-0,452	0,622	0,245	0,815	0,205
3	<i>insys</i>	-0,409	0,438	-0,430	0,324	0,167	0,192	0,185
4	<i>prdia</i>	-0,445	-0,555	-0,419	0,378	0,198	0,308	0,176
5	<i>papul</i>	-0,423	-0,330	-0,408	0,220	0,179	0,109	0,166
6	<i>pvent</i>	-0,474	-0,773	-0,437	0,531	0,224	0,598	0,191
7	<i>reput</i>	2,439	-0,105	2,447	0,156	5,949	0,011	5,987
<i>cliniques</i>		0,000	0,000	0,000	3,816	7,000	7,000	7,000

Les 2 variables sont parfaitement indépendantes puisque les sommes des densités de transformation sont nulles et que les sommes des carrés valent $n = 7$ ($7 =$ le nombre de facteurs linéaires, le nombre de *ddl* et la trace de la matrice de régression).

Les densités de V_t et de V_c des 6 premiers paramètres sont négatives et la 7^{ème} est positive, hautement significative (en rouge). Il y a là une symétrie qui confirme la parfaite liaison des variables.

Les densités des résidus V_r sont beaucoup plus contrastées. Le 1^{er} paramètre (*frcar*) a une densité hautement significative ($V_r = 2,229$ 'en rouge'). Le 3^{ème} paramètre (*insys*) est positif ($V_r = 0,438$). Et les 5 autres paramètres cliniques sont tous négatifs, affichant des valeurs moins uniformes que celles constatées dans V_t et V_c .

4.2.2 Les paramètres de base

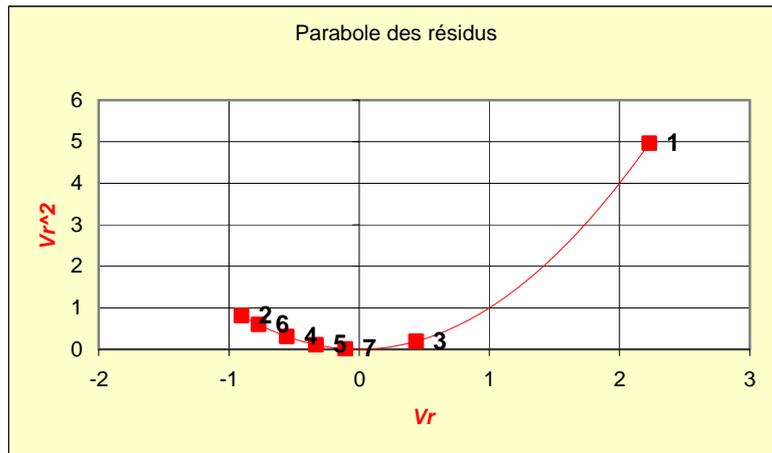
Les paramètres de base, directement calculés par la Macro, montrent que les 7 paramètres cliniques sont entièrement contrôlés (7 sur 7).

Avec un coefficient de corrélation $r = 0,999$ hautement significatif, un coefficient de détermination $r^2 = 0,998$. L'hypothèse retenue est celle d'une liaison stochastique à 95,3 %. La liaison stochastique des 2 sous-ensembles du corpus est parfaitement établie et entièrement contrôlée.

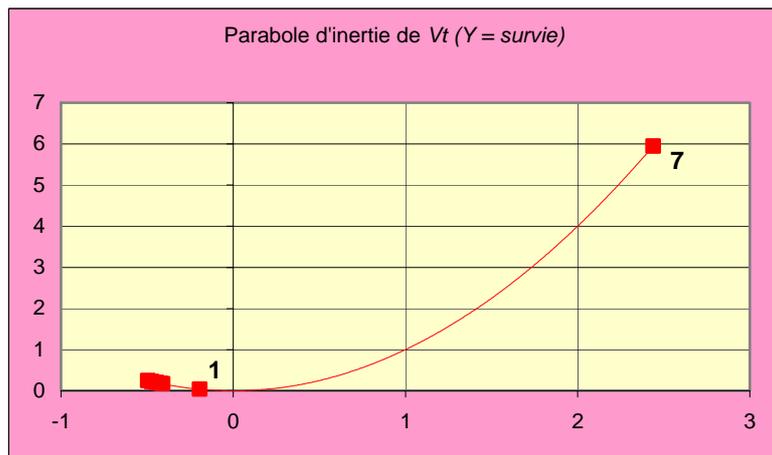
4.2.3 Les paraboles d'inertie

Regardons successivement les paraboles des densités de la transformation spectrale, puis le spectre des distances dt . Tout converge pour montrer l'unité du corpus et la pertinence de l'analyse statistique qu'on contrôle, qui se contrôle et qui contrôle tout.

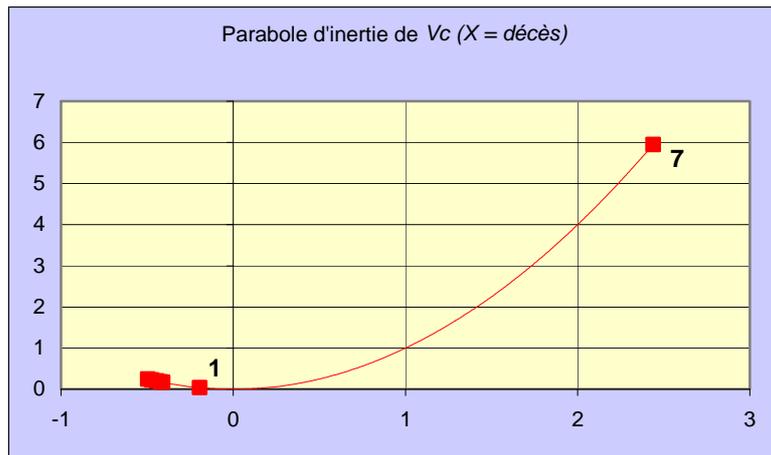
a) la parabole d'inertie des résidus V_r (survie / décès)



b) la parabole d'inertie des densités spectrales de V_t (= Y « survie »)



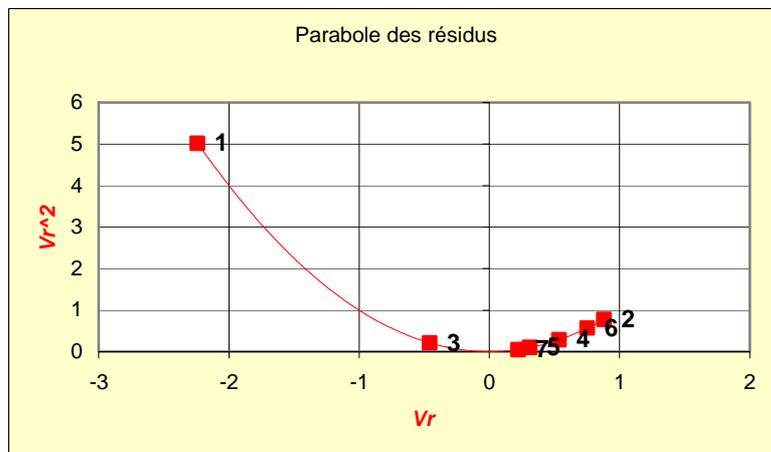
c) parabole d'inertie des densités spectrales de V_c (= X « décès »)



Les 2 paraboles d'inertie de X ($Vc = \text{décès}$) et de Y ($Vt = \text{survie}$) montrent, si besoin était, combien les 2 variables sont corrélées ($r = 0,999$). Les valeurs spectrales sont voisines.

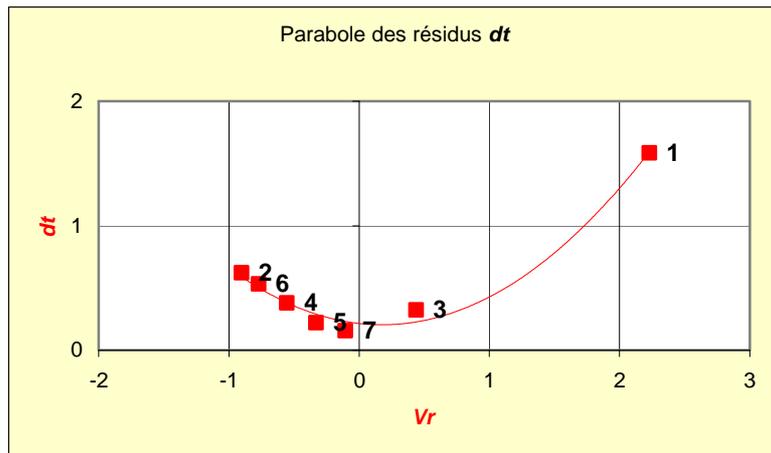
d) parabole d'inertie des résidus Vr ($\text{décès} / \text{survie}$)

Sachant que $Vt = Vc$ lorsqu'on inverse les variables, on imagine sans peine que les paraboles des résidus sont autant dire symétriques :



Il suffit de comparer les 2 paraboles pour s'en convaincre. C'est une symétrie quasi parfaite.

e) la parabole d'inertie de la distance dt ($= khi-2$) des résidus



La parabole d'inertie de la distance dt (khi-2 du spectre) montre à quel point les résidus jouent un rôle important dans l'analyse des phénomènes et dans la manière de décrypter les données.

On y voit, à l'extrémité de la branche montante de gauche, 3 paramètres sur la parabole, les n° 2, 6 et 4, (*incar*, *pvent* et *prdia*) et, à l'extrémité de la branche montante de droite, la variable n° 1 (*frcar*). Puis, les 3 variables restantes tendant vers le centre de gravité, la n° 5 et la n° 7 (*papul* et *repul*) en dessous de la parabole et la n° 3 (*insys*) au-dessus.

Dans le cas contraire, la parabole serait symétrique, mais les décalages se feraient dans le même sens : le 3 serait à l'intérieur et le 5 et le 7 à l'extérieur.

Remarque: la somme des carrés des valeurs de dt est toujours égale à $n/2$:

$$\sum (dt)^2 = \frac{n}{2}$$

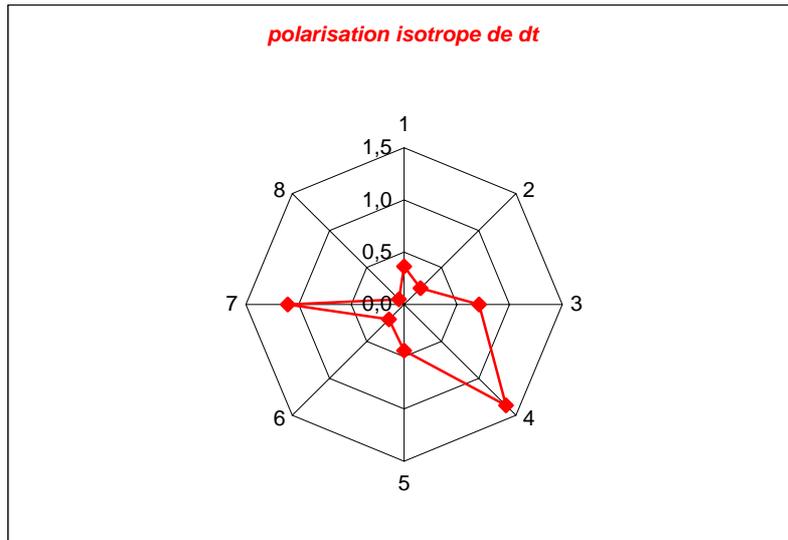
Le χ^2 ayant un caractère contraignant, il s'ensuit que les distances dt sont hautement significatives.

Corollaire :

- 1) $\sum(Vt)^2 = n, \sum(Vr)^2 = n, \sum(Vc)^2 = n$
- 2) $\sum(dt)^2 = n/2.$

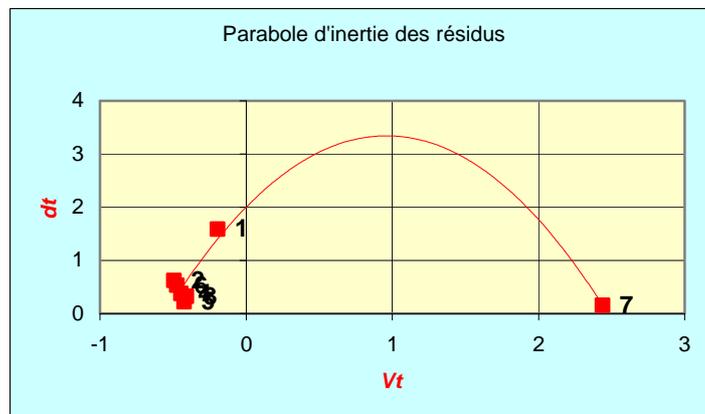
La vérification du corollaire de la transformation spectrale est immédiate.

Les vecteurs représentatifs des distances dt sont isotropes et hautement représentatifs du phénomène décrit.

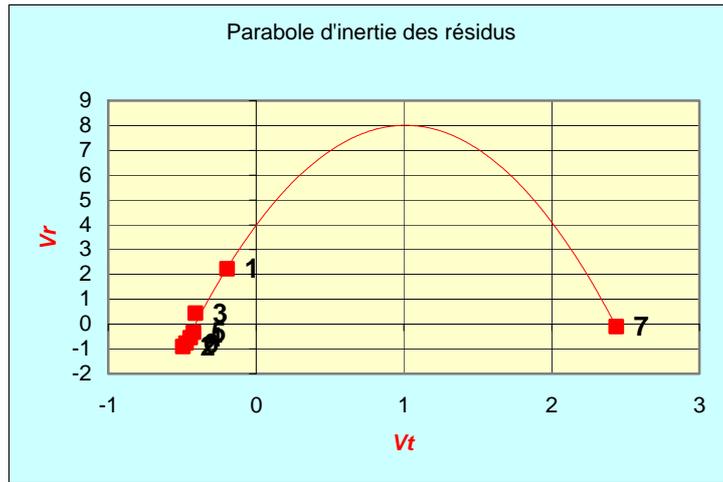


f) la parabole d'inertie des densités spectrales de dt ($= khi-2$) de Vt et de Vc

La parabole inversée des distances dt montre clairement la réalité des faits : 5 paramètres cliniques parfaitement groupés et 2 paramètres, 1 et 7, « excentrés » ou « déviants ».



La valeur du khi-2 ($dt = 3,816$) montre que la distribution est à 80 %. Il y a donc 2 facteurs sur 7 qui échappent au contrôle : ce sont les paramètres critiques 1 et 7 (*frcar* et *repu*) de la fréquence cardiaque et de la résistance pulmonaire qui accusent un résidu hautement significatif ($Vr = 2,229$) pour le paramètre 1 (*frcar*) ou une densité de base elle aussi hautement significative ($Vt = 2,439$ et $Vc = 2,447$) pour le 7^{ème} (*repu*).

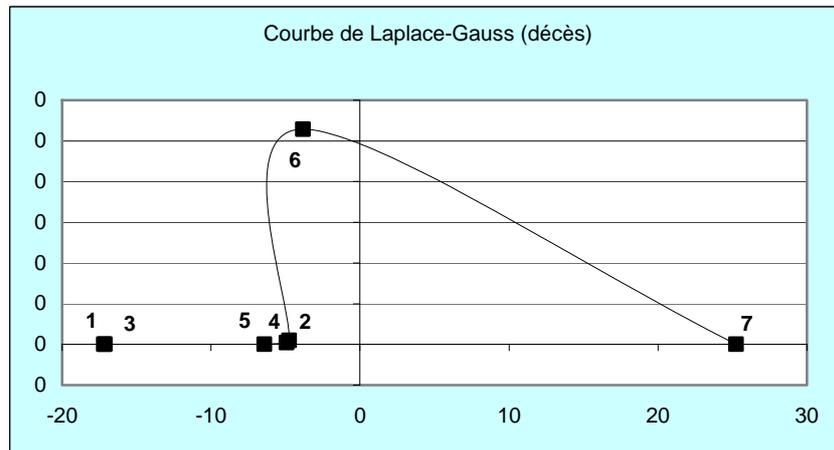


En résumé, on voit que les paraboles de V_t et de V_c véhiculent l'information de base émise par les « *probas* », alors que les paraboles d'inertie de V_r et de dt véhiculent une quantité d'informations que les spectres mettent clairement en évidence.

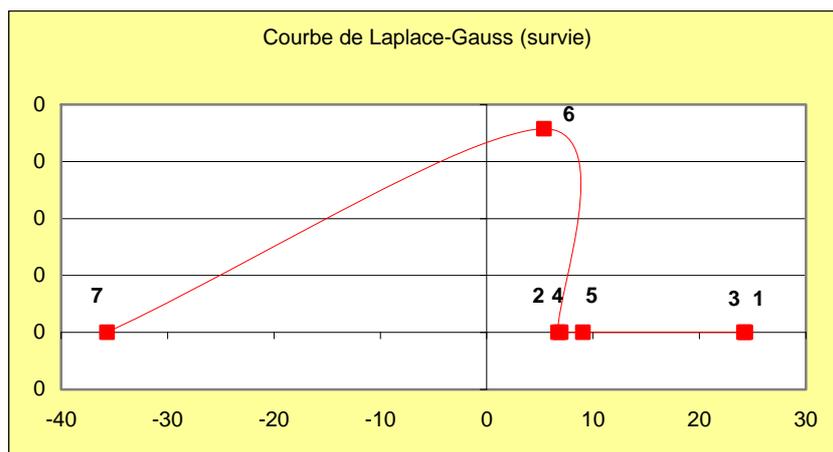
g) les courbes en cloche de Laplace-Gauss de V_t et de V_c

Les cloches de Laplace-Gauss montrent la complémentarité des valeurs de V_t et de V_c , en donnant une image inversée lorsqu'on inverse les paramètres cliniques de « *survie* » et de « *décès* » :

- la cloche des densités des « *décès* » :



- la cloche des densités des « *survie* » :



Le tout se reflète dans la partie et la partie se reflète dans le tout.

5. La régression linéaire simple

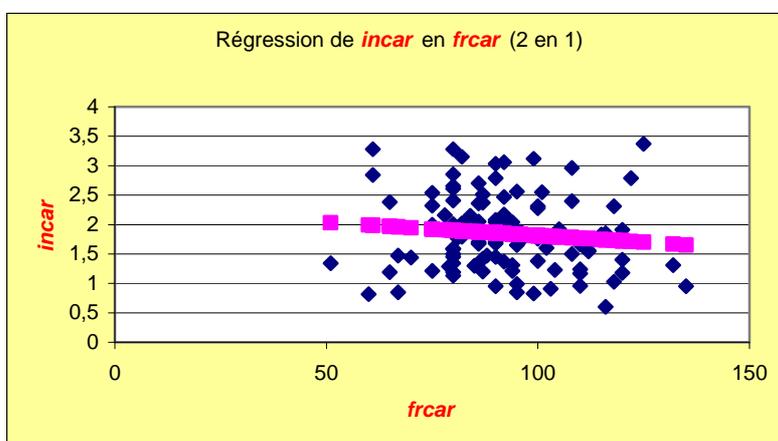
La régression linéaire multiple n'aurait aucun sens dans le cas présent. C'est une évidence. (Voir Chap. 5)

La régression linéaire simple, en revanche, permet de mesurer l'impact des paramètres cliniques appariés en vertu des *liaisons stochastiques* (Voir Chap. 5).

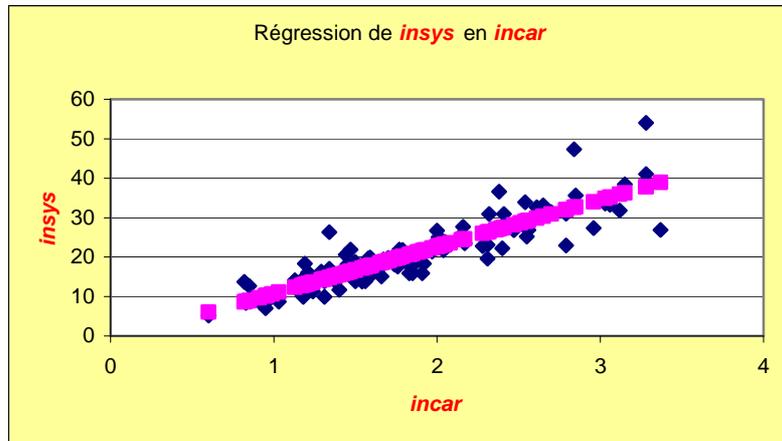
5.1 Images et degrés de liaison stochastique

En fonction des taux de corrélation, on sélectionne les variables appariées, on reporte les valeurs dans la feuille d'estimation et on va directement aux images qui préfigurent *les hypothèses des relations stochastiques* :

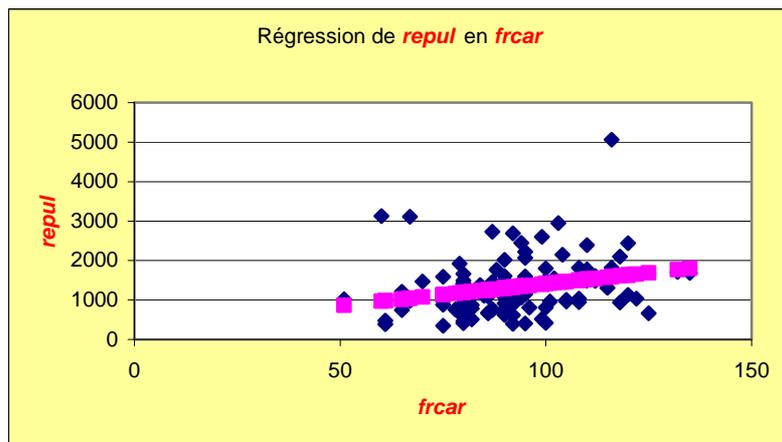
a) Avec une corrélation $r = -0,112$ les variables n° 1 et n° 2 (*frcar* et *incar*) de la *fréquence cardiaque* et l'*index cardiaque* accusent une absence de liaison.



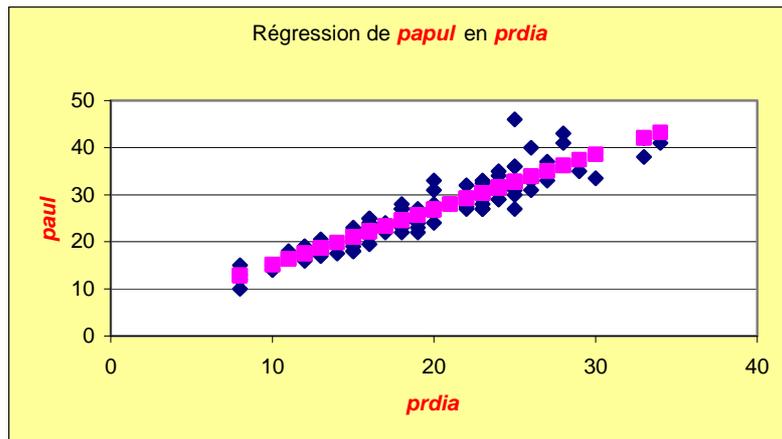
b) Avec une corrélation $r = 0,887$ les variables n° 2 et n° 3 (*incard* et *insys*) de l'*index cardiaque* et de l'*index systolique* affichent une liaison à 54%. Soit 54 individus liés et 47 déliés par cette relation.



c) Avec une corrélation $r = 0,247$ les variables n° 1 et n° 7 (*frcar* et *repul*) de la *fréquence cardiaque* et de la *résistance pulmonaire* accusent une résistance de 14%. Soit 14 individus déliés et 87 liés.



d) Avec une corrélation $r = 0,928$ les variables n° 4 et n° 5 (*prdia* et *papul*) de la *pression diastolique* et de la *pression artérielle pulmonaire* accusent une liaison de 63%. Soit 63 individus liés et 38 déliés.



C'est en connaissance de cause que l'analyse factorielle discriminante peut se poursuivre. Le coefficient de corrélation a fixé le cap de la *liaison stochastique* (le sens et le degré).

5.2 Sélection des paramètres d'analyse

Pour une analyse factorielle exhaustive, on fixera les paramètres en fonction des densités de transformation, des distances et aussi en fonction de la moyenne du « critère » (la variable expliquée Y).

Prenons quelques exemples pour essayer d'en comprendre les mécanismes et voir comment on isole les cas (facteurs, vecteurs ou individus).

a) Comparaison des paramètres cliniques 1 et 2 (*frcar* et *incar*)

1. 1 cas de non liaison et 100 cas de liaison
 2. Avec un $V_r > 0$ on a 48 cas de survie
 3. Avec un $V_r > 0$ et un $Y \geq \bar{Y}$, on a 45% de cas de survie
 4. Avec un $V_r > 0$ et un $Y \leq \bar{Y}$, on a 2 cas de décès confirmés
 5. Avec un $V_r > 0$ et un $dt \geq 1$, on a 6 cas de survie et 2 cas de décès
 6. Avec un $V_r > 0$ et un $V_t > 0$, on compte 45 cas de survie
 7. Avec un $V_r < 0$, on compte 50 cas de décès
- Et ainsi de suite.

b) Comparaison des paramètres cliniques 2 et 3 (*incar* et *insys*)

1. 54 cas sont liés et 47 résistent
 2. Avec un $V_r > 0$ et un $Y \geq \bar{Y} = 20,816$, on compte 26 cas de survie
 3. Avec un $V_r > 0$ et un $Y \leq \bar{Y} = 20,816$, on compte 20 cas de décès
 4. Avec un $V_r > 0$ et un $V_t > 0$, on compte 25 cas de survie
 5. Avec un $V_r > 0$ et un $V_t < 0$, on compte 20 cas de décès
 6. Avec un $V_r > 0$ et un $V_c > 0$, on compte 21 cas de survie
 7. Avec un $V_r > 0$ et un $V_c < 0$, on compte 24 cas de décès
- Et ainsi de suite.

c) *Comparaison des paramètres cliniques 1 et 7 (frcar et repul)*

2. 3 cas de non liaison et 98 cas de liaison
 3. Avec un $V_r > 0$ on a 43 cas de décès
 4. Avec un $V_r > 0$ et un $Y \geq \bar{Y}$, on a 39 cas de décès
 5. Avec un $V_r > 0$ et un $V_t > 0$, on a 37 cas de décès
 6. Avec un $V_r > 0$ et un $V_c < 0$, on compte 21 cas de décès
 7. Avec un $Y \geq \bar{Y} = 1324$, on compte 44 cas de décès
 8. Avec un $Y \leq \bar{Y} = 1324$, on compte 50 cas de survie
- Et ainsi de suite.

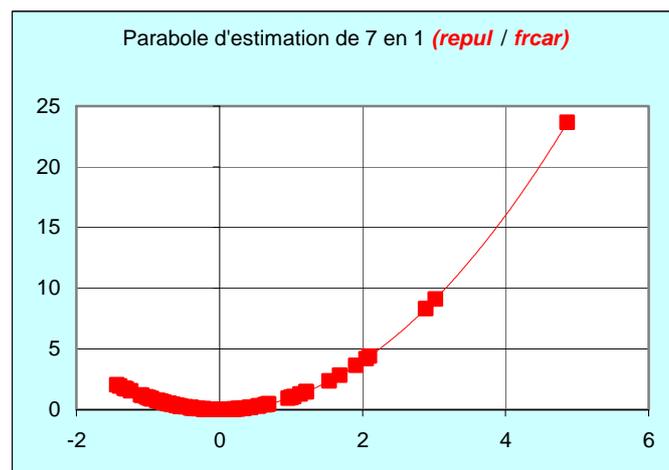
d) *Comparaison des paramètres cliniques 4 et 5 (prdia et papul)*

1. 63 cas de liaison et 38 cas de non liaison
 2. Avec un $V_r > 1,96$ on a 4 cas de décès
 3. Avec un $-1,96 \leq V_r \leq +1,96$ on a 50 cas de survie et 46 décès
 4. Avec un $V_r \leq -1,96$ on a 1 décès
- Et ainsi de suite.

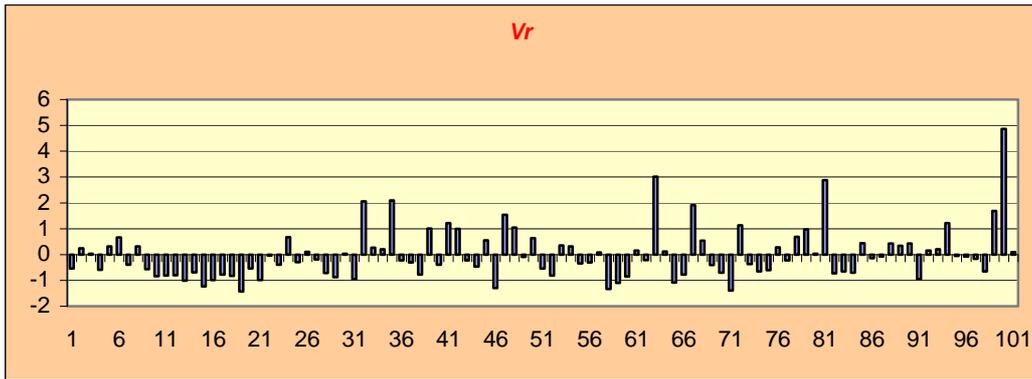
La statistique mesure en permanence le tout et la partie pour en dévoiler les liaisons.

5.3 Les graphiques

La parabole d'inertie des résidus de l'estimation de 7 en 1 (*repul / frcar*), du rapport de la résistance pulmonaire à la fréquence cardiaque, montre clairement l'ampleur du phénomène dans toute son étendue, pour les 101 de « décès » et de « survie ».



L'histogramme donne clairement la position de chaque individu.



On recherchera toujours le graphique le plus performant, celui qui permet de visualiser au mieux l'ampleur du phénomène analysé.

6. Pronostics, Prévisions et Décisions

En guise de conclusion, on voit que pronostics, prévisions, estimations et décisions sont mis à disposition par la Macro qui traite immédiatement les données introduites dans la structure.

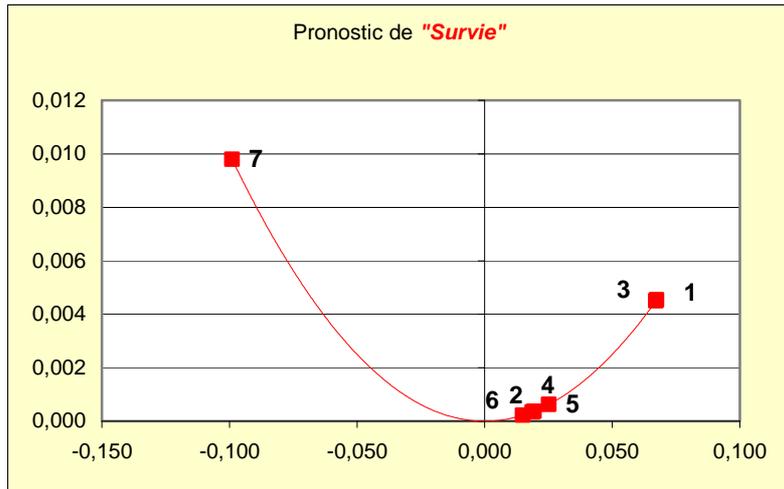
Exercice pratique. Pour pronostiquer les chances de « *Survie* » ou de « *Décès* » d'un patient :

- 1) introduire les mesures des paramètres cliniques dans la règle
- 2) analyser les calculs et les graphiques d'appréciation
- 3) approfondir les estimations et les spectres
- 4) faire les analyses discriminantes (ACP, AFD, ASR)
- 5) analyser les spectres
- 6) vérifier la cohérence des résultats

Pour ce faire, on dispose de la Macro et des mesures statistiques pour traiter les données, et des échelles de référence pour interpréter les résultats.

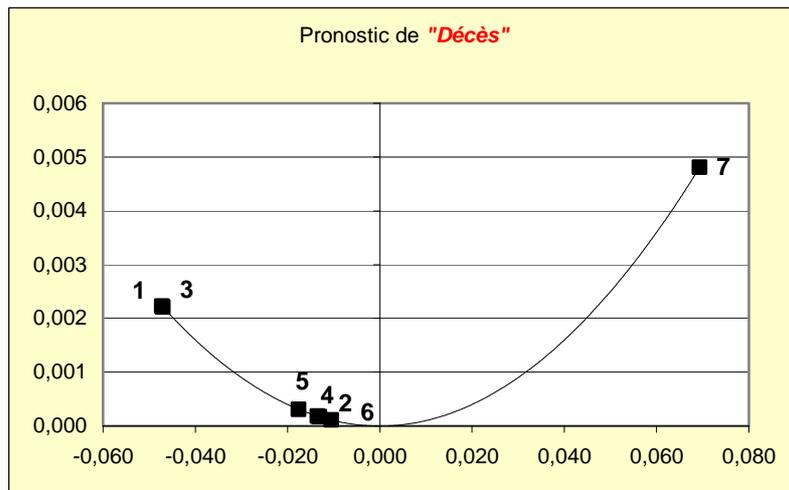
6.1.1 Règle et mesures moyennes pour le pronostic de « *Survie* »

	150857,32	9308	186,42	2102,4	1945	2626	959,5	133730
	<i>p</i>	0,062	0,001	0,014	0,013	0,017	0,006	0,886
	<i>q</i>	0,938	0,999	0,986	0,987	0,983	0,994	0,114
	Fréq	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
<i>moyenne</i>	1006,45	88,34	2,31	26,75	16,50	22,84	8,33	841,38
<i>densité</i>	5,797	3,438	0,953	3,422	0,985	1,282	0,765	-5,048
<i>moyenne</i>	0,116	0,067	0,019	0,067	0,019	0,025	0,015	-0,099
<i>Moy^2</i>		0,005	0,000	0,005	0,000	0,001	0,000	0,010
		1	2	3	4	5	6	7



6.1.2 Règle et mesures moyennes pour le pronostic de «Décès »

	150857,32	9308	186,42	2102,4	1945	2626	959,5	133730
<i>p</i>		0,062	0,001	0,014	0,013	0,017	0,006	0,886
<i>q</i>		0,938	0,999	0,986	0,987	0,983	0,994	0,114
	Fréq	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
<i>moyenne</i>	1971,27	95,90	1,39	15,00	21,96	29,10	10,65	1797,27
<i>densité</i>	-4,061	-2,408	-0,668	-2,397	-0,690	-0,898	-0,536	3,536
<i>moyenne</i>	-0,080	-0,047	-0,013	-0,047	-0,014	-0,018	-0,011	0,069
<i>moy^2</i>		0,002	0,000	0,002	0,000	0,000	0,000	0,005
		1	2	3	4	5	6	7



La Macro traite automatiquement les données et fixe un diagnostic que le médecin peut interpréter à bon droit et en parfaite connaissance de cause.

On dispose d'une règle de mesure et de 2 échelles d'estimation des risques de « *survie* » ou de « *décès* » pour établir un pronostic fiable :

6.2.1 Échelle des valeurs moyennes des paramètres cliniques

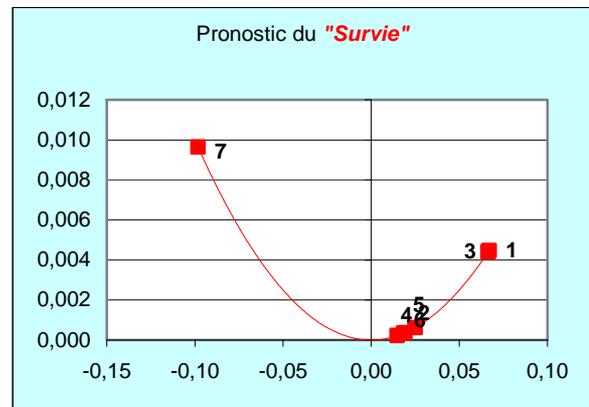
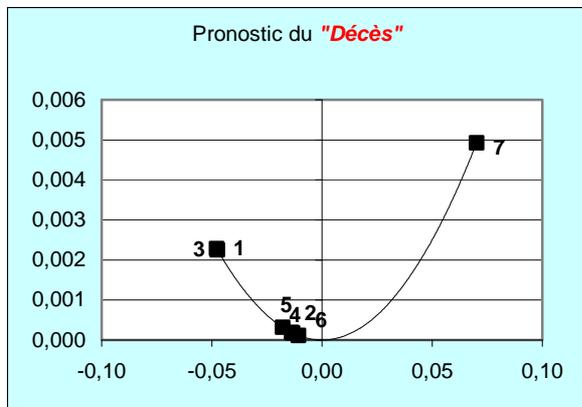
moyennes	Total	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
SURVIE	1006,45	88,34	2,31	26,75	16,50	22,84	8,33	841,38
DÉCÈS	1971,27	95,90	1,39	15,00	21,96	29,10	10,65	1797,27
paramètres		1	2	3	4	5	5	7

On pourrait fixer les intervalles de variation autour de la moyenne à $m \pm 2z$ pour voir si les mesures effectuées sont dans la fourchette. Mais en visant directement les densités moyennes on y gagne en temps, en précision et en interprétation.

6.2.2 Échelle des moyennes de densité

densités	Moy	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
SURVIE	0,828	3,438	0,953	3,422	0,985	1,282	0,765	-5,048
DÉCÈS	-0,580	-2,408	-0,668	-2,397	-0,690	-0,898	-0,536	3,536
paramètres		1	2	3	4	5	5	7

Les « *pesées* » faites avec la règle « *se prononcent* » sur les chances de « *survie* » ou de « *décès* » d'un patient.



Les deux paraboles (à échelle comparative) montrent que les paramètres 1 (*frcar*), 3 (*insys*) et 7 (*repul*) sont des paramètres déterminants de tout diagnostic, même si l'on ne peut pas négliger les 4 autres qui tournent sensiblement autour de la moyenne.

6.2.3 Vérifications

La vérification est immédiate. Il suffit de reprendre les mesures effectuées sur les patients, de les projeter dans la règle et de lire les résultats à la lumière des « critères normaux ». Vérification qui est un pur exercice d'école, puisque les résultats sont déjà connus et donnés par la TDR. Au-delà de la tautologie, il s'agit de se familiariser avec un exercice d'école pour mettre la statistique à la portée de tous.

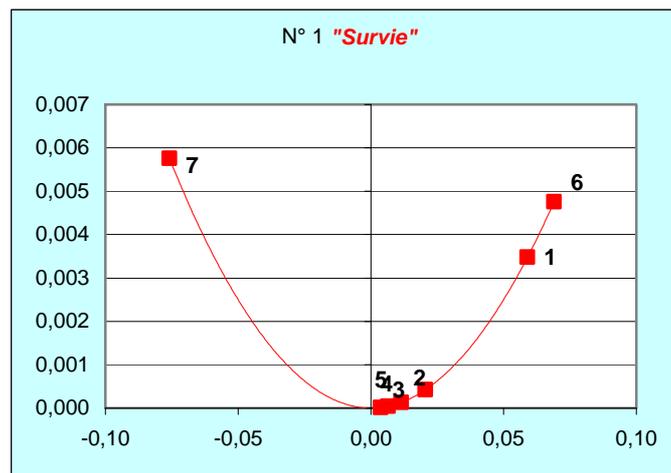
Exemples :

1. *Le cas n° 1 est un cas de « survie ».* Quels sont les points faibles et les points forts ?

La règle donne les « densités » suivantes :

<i>densités</i>	Moy	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
SURVIE	0,828	3,438	0,953	3,422	0,985	1,282	0,765	-5,048
mesures	0,686	3,008	0,332	1,049	0,582	0,187	3,519	-3,871
paramètres		1	2	3	4	5	5	7

Observations. La moyenne est légèrement inférieure à la moyenne générale et les valeurs algébriques vont toutes dans le bon sens. Les critères 1, 2, 3, 4, 5 et 7 sont tous inférieurs à la « norme », mais certains accusent une faiblesse comme le 2 (*incar*), le 3 (*insys*), et surtout le 4 (*papul*). Mais on remarque surtout l'inversion du critère 5 (*pvent*) qui est 5 fois supérieur à la norme.



La ligne des points faibles et des points forts est nettement cernée.

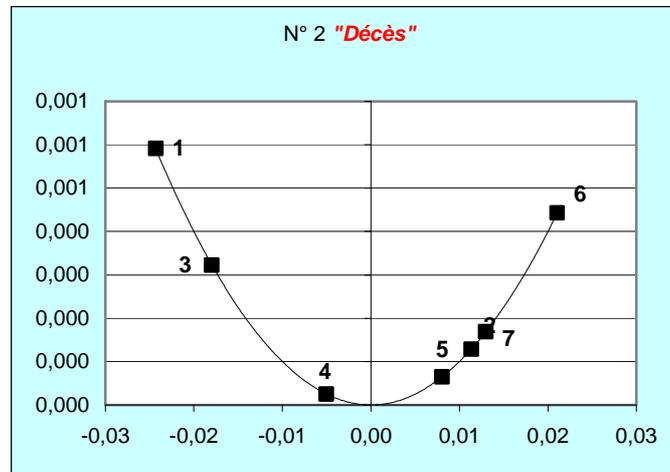
2. *Le cas n° 2 est un cas de « décès ».* Quels sont les points faibles et les points forts ?

La règle donne les « densités » suivantes :

<i>densités</i>	Moy	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
DÉCÈS	-0,580	-2,408	-0,668	-2,397	-0,690	-0,898	-0,536	3,536
mesures	0,045	-1,240	-0,256	-0,916	0,579	0,411	1,073	0,663
paramètres		1	2	3	4	5	5	7

Observations. Une moyenne générale positive, donc tout n'est pas « négatif » *a priori*. En effet, les critères 4 (*prdia*), 5 (*papul*) et 6 (*pvent*) sont positifs, comme dans les cas de

« survie ». Ce sont les points forts. En revanche, tous les autres critères vont dans le mauvais sens, le 1 (*frcar*), le 2 (*incar*), le 3 (*insys*) et le 7 (*repul*), même si l'ensemble reste « modéré ». Quelles en sont les causes ? Voilà le problème qu'il faut résoudre.



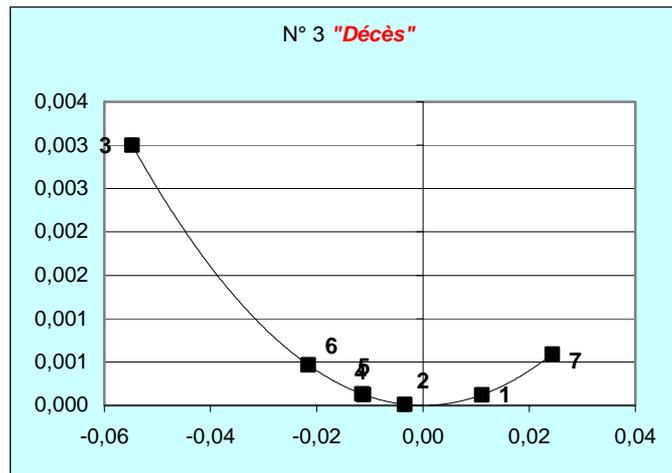
L'image focalise les points forts et les points faibles.

3. Le cas n° 3 est encore un cas de « décès ». Quels sont les points faibles et les points forts ?

La règle donne les « densités » suivantes :

densités	Moy	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
DÉCÈS	-0,580	-2,408	-0,668	-2,397	-0,690	-0,898	-0,536	3,536
mesures	-0,489	0,565	-0,586	-2,793	-0,176	-0,570	-1,102	1,242
paramètres		1	2	3	4	5	5	7

Observations. Une moyenne générale négative, ce qui est en soi mauvais signe. Excepté le critère 1 (*frcar*) qui va dans le bon sens puisqu'il est positif comme dans els cas de « survie », tous les autres critères penchent du mauvais côté, notamment le 5 (*pvent*) qui accuse une grande faiblesse.



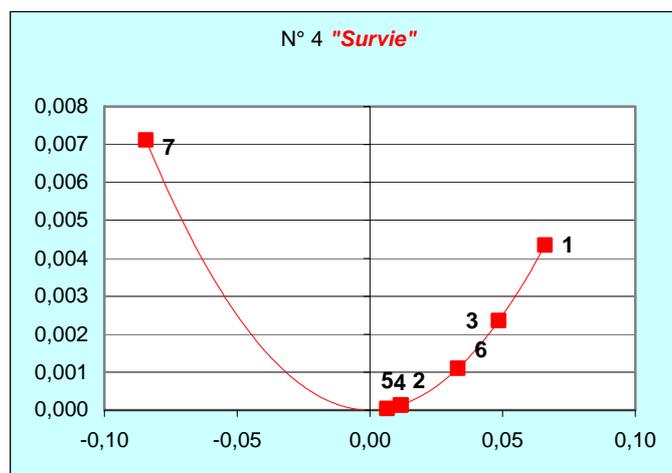
On peut comparer à travers les images les deux cas de « décès ».

4. Le cas n° 4 est de nouveau un cas de « survie ». Quels sont les points faibles et les points forts ?

La règle donne les « densités » suivantes :

<i>densités</i>	Moy	<i>frcar</i>	<i>incar</i>	<i>insys</i>	<i>prdia</i>	<i>papul</i>	<i>pvent</i>	<i>repul</i>
SURVIE	0,828	3,438	0,953	3,422	0,985	1,282	0,765	-5,048
mesures	0,678	3,364	0,599	2,477	0,588	0,333	1,688	-4,303
paramètres		1	2	3	4	5	5	7

Observations. La moyenne est légèrement inférieure à la moyenne générale. Le critère 1 (*frcar*) est bon. Les critères 2 (*incar*), 3 (*insys*), 4 (*prdia*), 5 (*papul*) et 7 (*repul*) accusent une faiblesse certes, mais ils vont dans le bon sens. En revanche, le critère 5 (*pvent*) est supérieur à la norme. Il y a là un point faible manifeste.



La parabole focalise la ligne des critères qui montre les points forts et les points faibles.

Et ainsi de suite...

Voilà comment on peut techniquement procéder pour formuler un pronostic. On pourrait certes reprendre tous les cas, mais on peut aussi bien se reporter directement aux données de la TDR et les analyser à la lumière des « normes » qui pronostiquent les chances de « décès » ou de « survie ».

*
* *

La Macro offre toutes les possibilités de contrôle et de vérification. Tout est toujours possible : lemmatisations, discriminations, ACP, AFD, ASR... On peut renouveler les mesures, ajuster les paramètres, filtrer les variables, calculer les densités, analyser les spectres, multiplier les graphiques. Tout est vérifié et tout est vérifiable. La statistique montre et démontre, dit et redit.

La statistique a la capacité de mesurer et de décrire les phénomènes et d'aider à les interpréter.

L'informatique a la puissance du calcul.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

7

La régression linéaire simple Traitement d'un échantillon Formules analytiques

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce septième chapitre, notre intention est de proposer un rappel de la construction et de l'étude de la droite de régression en présence d'un échantillon (sous-ensemble de la population étudiée). Nous appliquerons ensuite la partie théorique à l'étude d'un exemple.

1. La régression linéaire

Deux paramètres, X et Y , ayant été observés, il peut alors être intéressant de savoir s'il est possible de faire passer une droite au milieu de ces observations afin d'établir une relation de type affine entre X et Y : $Y' = b + aX$. Une fois le modèle linéaire établi, nous verrons comment valider ce modèle : la droite calculée ne passant généralement pas par tous les points expérimentaux, nous vérifierons que cet écart est bien dû au hasard. Enfin nous ferons une exploitation du modèle avec la détermination de l'intervalle de confiance et de la plus petite valeur mesurable.

1.1 Etablissement du modèle linéaire

1. Choix du modèle linéaire

Avant de poser le modèle de régression on doit faire une étude descriptive simple des données : si le nuage de points obtenu en traçant le graphique des observations du couple (X, Y) est assez « allongé », on peut envisager une régression linéaire.

Pour renforcer l'idée que la relation entre les variables va être linéaire, le coefficient de corrélation linéaire empirique doit être proche de 1 en valeur absolue.

Rappel : Coefficient de corrélation linéaire de Pearson :

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$$

avec $C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, $\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ et $\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$.

2. Le modèle

Notons :

- y_i : les observations de Y
- y'_i : l'estimation de y_i obtenue par la droite de régression ($y'_i = b + ax_i$)
- e_i : l'écart entre les observations et le modèle ($= y_i - y'_i$)

Pour que la droite passe au plus près des observations, pour qu'elle s'ajuste au mieux à l'ensemble des points observés, il faut que ces écarts soient aussi petits que possible. La droite de régression doit donc être la droite qui minimise l'ensemble de ces écarts. Comme les écarts peuvent être positifs ou négatifs, le problème de l'influence du signe sera éliminé en considérant le carré des écarts.

Notons :

- SCE (Somme des Carrés des Ecarts) = $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - y'_i)^2$.

Il s'agit donc de minimiser

$$SCE = \sum_{i=1}^n [y_i - (b + ax_i)]^2,$$

c'est-à-dire qu'il faut trouver les paramètres a et b tels que la somme des carrés des écarts soit minimale.

- SCE sera minimale si les deux dérivées partielles par rapport à a et b sont nulles :

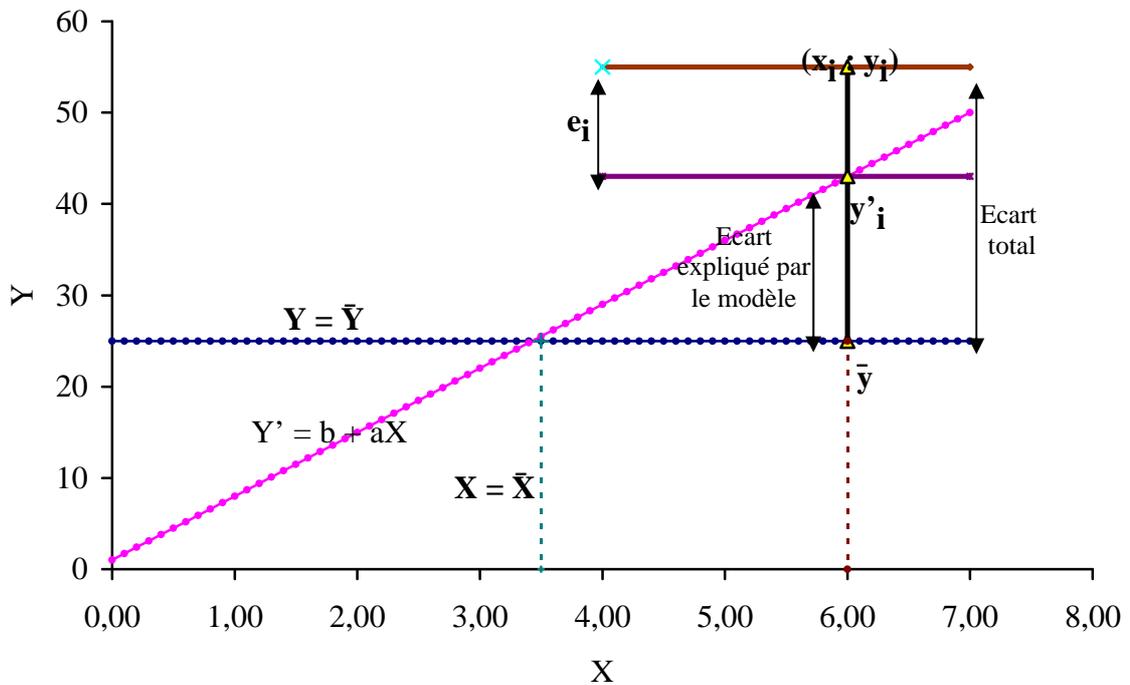
$$\frac{\partial SCE}{\partial a} = 0 \text{ et } \frac{\partial SCE}{\partial b} = 0.$$

La dérivation partielle donne

$$\begin{cases} b \sum_{i=1}^n x_i + a \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \\ nb + a \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \end{cases}$$

Il faut ici résoudre un système de deux équations à deux inconnues. Le résultat en est :

$$b = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad \text{et} \quad a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} .$$



Ce graphique représente l'ensemble des notations données plus haut, mais aussi montre comment se construit l'équation de l'analyse de la variance en régression linéaire simple donnée au paragraphe 3 du chapitre 3 :

$$\sum (y_i - \bar{y})^2 = \sum (y_i - y'_i)^2 + \sum (y'_i - \bar{y})^2$$

soit encore :

$$SCT = SCR + SCE$$

où SCT est la somme des carrés totale (à (n-1) ddl), SCR est la somme des carrés des résidus (à (n-2) ddl) et SCE est la somme des carrés des écarts expliqués par le modèle (à 1 ddl).

1.2 Validation du modèle linéaire

Une fois la relation linéaire entre X et Y établie, il reste encore à valider le modèle trouvé. Pour cela il faut faire une étude des résidus.

La relation est $Y = b + aX$ avec les paramètres a et b qui ont été calculés au paragraphe précédent. Considérons l'hypothèse nulle H_0 selon laquelle le modèle est bien cette droite. Sous cette hypothèse, si on répète pour chaque valeur de X chacune des mesures Y un grand nombre de fois, les écarts vont avoir une distribution statistique car sous H_0 , la variation des y_i autour des y'_i est uniquement due au hasard. En fait, les e_i ont une distribution normale centrée, d'écart type $\sigma : N(0 ; \sigma)$. Les erreurs sont indépendantes et σ est constant (indépendant de x).

L'écart type reste inconnu puisqu'on ne répète pas chaque mesure une infinité de fois. On utilisera donc un estimateur de σ .

On définit une Erreur Moyenne d'Ajustement (EMA) notée $S(y,x)$ qui est une estimation de σ . C'est l'erreur mathématique due au modèle.

$$S(y,x) = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - y'_i)^2}{n-2}}$$

Remarque : d'une manière générale, l'estimateur d'un écart type sur un **échantillon**

est toujours donné par
$$S = \sqrt{\frac{\sum_{i=1}^n (y_i - y'_i)^2}{ddl}}$$

Le ddl (nombre de degrés de liberté) est en fait égal à $[n - \text{nombre de paramètres estimés}]$: en effet pour une régression linéaire simple, il y a deux paramètres (a et b) à estimer ce qui donne un ddl de $[n - 2]$; pour une régression linéaire multiple de la forme $Y = b + a_1X_1 + a_2X_2 + \dots + a_pX_p$ le ddl sera égal à $[n - (p + 1)]$ puisqu'il y a $p + 1$ paramètres ($b ; a_1 ; a_2 ; \dots ; a_p$).

Sous H_0 , nous avons donc

$$S(y,x) = \sqrt{\frac{\sum_{i=1}^n y_i^2 - b \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i y_i}{n-2}}$$

Les résidus suivent donc une loi normale $N(0, S(y,x))$. Comme il n'existe une table que pour la loi normale centrée réduite $LG(0,1)$, il faut maintenant considérer les écarts réduits (ou résidus Studentisés) : $\frac{e_i}{S(y,x)}$ qui suivent une $LG(0,1)$ sous H_0 . Ainsi $\frac{e_i}{S(y,x)}$ a 95% de chances de se trouver entre les bornes $-1,96$ et $+1,96$.

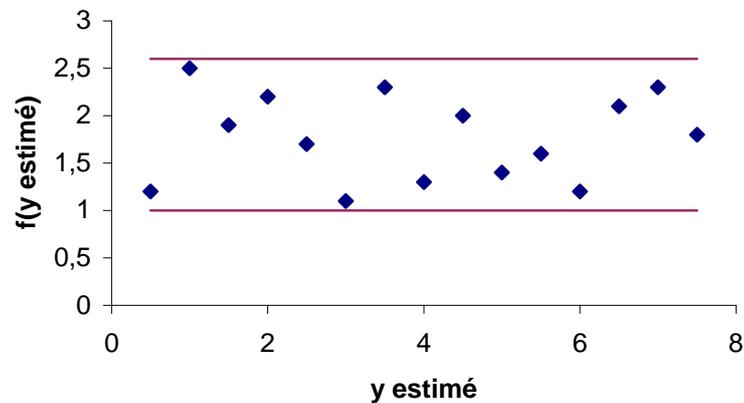
La méthode de validation est alors la suivante.

1. Tracer $f(y'_i) = \frac{e_i}{S(y,x)}$

On réalise le nuage de points avec en abscisse les y'_i (parfois y_i) et en ordonnée les $\frac{e_i}{S(y,x)}$. C'est le graphique des résidus. Il sert à contrôler les différentes hypothèses du modèle.

2. Contrôle visuel de l'indépendance des résidus et de l'homoscédasticité

Sur le graphique, les résidus doivent être éparpillés sans forme apparente (signe d'indépendance des erreurs) selon une bande horizontale (signe d'une variance constante et donc d'une bonne homoscélasticité) comme montré dans le graphique ci-dessous.



3. Contrôle de la normalité des erreurs

Nous savons que sous H_0 les erreurs e_i sont normales et que les $\frac{e_i}{S(y,x)}$ sont normales réduites. Donc si 95% des $\frac{e_i}{S(y,x)}$ se trouvent dans l'intervalle $[-1,96 ; +1,96]$, alors on retient H_0 , le modèle est validé.

S'il existe plus de 5% des valeurs à l'extérieur de l'intervalle, on rejette H_0 , le modèle n'est pas bon.

Remarque : lorsque l'on est en présence de petits échantillons ($n < 30$), les écarts réduits suivent une loi de Student à $n-2$ ddl. Il faut donc, pour $\alpha = 5\%$, que les $\frac{e_i}{S(y,x)}$ se trouvent dans l'intervalle $[-t(\alpha/2 ; ddl) ; +t(\alpha/2 ; ddl)]$, où la valeur critique $t(\alpha/2 ; ddl)$ se lit dans la table de la loi de Student.

4. Inférence finale

L'estimateur de l'écart type des résidus permet aussi de calculer l'écart type des paramètres a et b :

$$\sigma(a) = \frac{S(y,x)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad \sigma(b) = S(y,x) \cdot \sqrt{\frac{\sum_{i=1}^n x_i^2 / n}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Ces deux écart types permettent de tester la nullité des coefficients de régression a et b.

Test de la pente : il permet de tester la significativité du modèle

Considérons les hypothèses à tester $H_0 : a = 0$ et $H_1 : a \neq 0$.

Le test sera fait avec un risque $\alpha = 5\%$ (c'est le risque de se tromper en rejetant à tort H_0).

La statistique de test utilisée est : $T = \frac{a - 0}{\sigma(a)}$ qui suit une loi de Student à $n-2$ ddl.

Règle de décision et conclusion :

- si $|T| > t_{\alpha, \text{ddl}=n-2}$, rejet de H_0 : on a bien une dépendance linéaire entre X et Y de pente a
- si $|T| < t_{\alpha, \text{ddl}=n-2}$, non rejet de H_0 : Y est constant et ne dépend pas de X

Test de l'ordonnée à l'origine : cela permet de simplifier le modèle

Considérons les hypothèses à tester $H_0 : b = 0$ et $H_1 : b \neq 0$.

Le test sera fait avec un risque $\alpha = 5\%$

La statistique de test utilisée est : $T = \frac{b - 0}{\sigma(b)}$ qui suit une loi de Student à $n-2$ ddl.

Règle de décision et conclusion :

- si $|T| > t_{\alpha, \text{ddl}=n-2}$, rejet de H_0 : on ne peut pas simplifier le modèle ; la droite a bien b comme ordonnée à l'origine et ne passe pas par 0
- si $|T| < t_{\alpha, \text{ddl}=n-2}$, non rejet de H_0 : on peut simplifier le modèle en prenant 0 comme ordonnée à l'origine.

Remarque : en remplaçant 0 par n'importe quelle autre valeur dans les statistiques de test, il est possible de faire d'autres comparaisons.

1.3 Exploitation du modèle linéaire validé : l'intervalle de confiance

D'après la méthode des résidus, à chaque x_i correspond une mesure y_i dont l'écart à la droite de régression suit une loi normale.

Si on se fixe un niveau de confiance $1 - \alpha$ (95%), l'intervalle de confiance est la portion de la gaussienne qui contient $(1 - \alpha)\%$ de toutes les valeurs possibles de y. Les deux bornes de l'intervalle de confiance se calculent alors avec la formule suivante

$$IC = b + ax_i \pm t_{\alpha, \text{ddl}=n-2} \cdot S(y, x) \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

où n est le nombre de couples (x_i, y_i) et \bar{x} est la moyenne des x_i .

D'après cette formule, on peut remarquer que plus x_i s'éloigne de \bar{x} , plus $(x_i - \bar{x})^2$ augmente et plus l'enveloppe s'éloigne de la droite, d'où une enveloppe (un intervalle de confiance autour de la droite) en double trompette.

Une fois cet intervalle de confiance calculé, pour toute observation supplémentaire d'un couple $(x_i ; y_i)$, si le point est situé à l'intérieur de l'intervalle de confiance, il sera considéré comme appartenant à la droite.

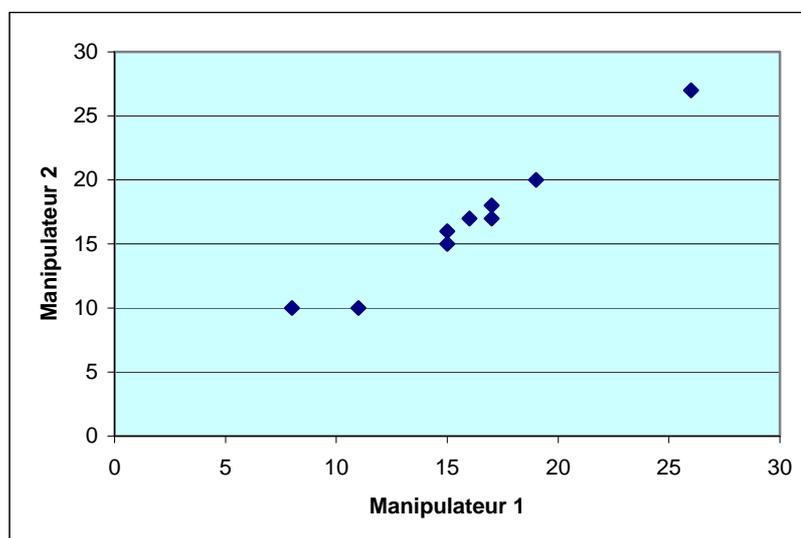
2. Etude d'un exemple

Dans ce paragraphe nous allons étudier l'exemple fictif suivant : pour tester la dextérité d'un nouveau manipulateur (manipulateur 2), il doit refaire tout les dosages de concentration en insuline (en $\mu\text{g/ml}$) de 9 échantillons différents. Ses mesures seront alors comparées à celles effectuées par un manipulateur expérimenté (manipulateur 1) sur les mêmes échantillons. Si le nouveau manipulateur est « bon », il devrait avoir statistiquement les mêmes résultats que le manipulateur 1. C'est ce que nous allons vérifier.

Les mesures des deux manipulateurs sont les suivantes :

Manipulateur 1	Manipulateur 2
11	10
15	15
19	20
16	17
17	17
8	10
26	27
17	18
15	16

Nous allons, avant de nous lancer dans des calculs, regarder comment se comporte le graphique de ces données :



On peut remarquer que les données ont une forme allongée, on va donc pouvoir essayer d'établir une relation linéaire entre les deux séries de mesures. Pour cela, on peut faire les calculs décrits au paragraphe 1 à l'aide des fonctions « simples » d'Excel (comme `COEFFICIENT.CORRELATION`, `COEFFICIENT.DETERMINATION`, `DROITEREG`, `ORDONNEE.ORIGINE`, `PENTE`, etc.) ou bien faire appel à l'outil « Utilitaire d'analyse » pour la « Régression linéaire » qui va fournir automatiquement l'ensemble des calculs statistiques dont nous avons besoin.

Après avoir fait fonctionner l'utilitaire d'analyse, nous obtenons l'ensemble des tableaux et graphiques suivants :

1) le premier tableau récapitule l'ensemble des coefficients

<i>Statistiques de la régression</i>	
Coefficient de détermination multiple	0,98579487
Coefficient de détermination R ²	0,97179152
Coefficient de détermination R ²	0,96776174
Erreur-type	0,92429108
Observations	9

Avec le graphique des données, les coefficients de détermination nous indiquent que nous pouvons continuer notre analyse. La valeur de l'erreur-type correspond à ce que nous avons appelé l'erreur moyenne d'ajustement.

2) le deuxième tableau donne toute l'analyse de variance. D'une façon plus générale voici ce qu'il donne comme résultats :

ANALYSE DE
VARIANCE

	<i>Degré de liberté</i>	<i>Somme des carrés</i>	<i>Moyenne des carrés</i>	<i>F</i>	<i>Valeur critique de f</i>
Régression	1	SCE	=SCE/1	=[SCE/1]/[SCR/(n-2)]	p-value
Résidus	n-2	SCR	=SCR/(n-2)		
Total	n-1	SCT			

Le test F permet de tester la significativité du modèle trouvé. Il teste l'hypothèse nulle H₀ : « Y ne dépend pas de X » contre l'alternative H₁ : « Y dépend de X à travers la relation Y = b + aX avec les a et b calculés ». La statistique F est donnée par la formule

$$F = \frac{SCE/1}{SCR/(n - 2)}$$

Pour un risque α donné (fixé à 5% par défaut dans l'utilitaire d'analyse),

- si F est supérieur à la valeur critique de F, on rejette H₀ et on accepte H₁ au risque α : Y dépend de X et Y = b + aX
- si F est inférieur à la valeur critique de F, on ne rejette pas H₀ au risque α : on ne peut pas dire que Y dépend de X.

La valeur critique de F se calcule alors avec la fonction INVERSE.LOI.F appliquée aux arguments (α ; 1 ; n-2).

ATTENTION : dans la sortie de l'utilitaire d'analyse, la valeur indiquée comme étant la valeur critique de F est en fait la « p-value », c'est-à-dire la probabilité sous H₀ d'observer la valeur F trouvée. La règle de décision et la conclusion du test à partir de la p-value sont les suivantes :

- si p-value < α : on rejette H₀ (la probabilité d'observer la valeur de F est très faible) et donc on accepte H₁ : Y dépend de X et Y = b + aX

- si $p\text{-value} > \alpha$: on ne rejette pas H_0 au risque α : on ne peut pas dire que Y dépende de X.

ANALYSE DE VARIANCE					
	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de f
Régression	1	206,019802	206,019802	241,152318	5,591460
Résidus	7	5,98019802	0,854314003		1,10939E-06
Total	8	212			

Plus concrètement, l'analyse de variance des données nous donne donc une valeur F de 241,152318 et une p-value de 1,10939E-06 qui est largement plus petite que le risque α de 5%. On peut donc rejeter l'hypothèse nulle selon laquelle Y ne dépend pas de X et accepter son alternative : Y dépend de X et $Y = 0,50825083 + 1,00990099X$ (les valeurs de la pente et de l'ordonnée à l'origine sont à lire dans le tableau suivant).

Remarque : la valeur critique de f vaut 5,59145974 et conduit à la même conclusion.

3) le troisième tableau donne tout ce qui concerne la pente et l'ordonnée à l'origine.

	Coefficients	Erreur-type	Statistique t	Probabilité
Constante	0,50825083	1,08518144	0,46835562	0,6537641
Manipulateur 1	1,00990099	0,06503289	15,5290797	1,1094E-06

	Limite inférieure pour seuil de confiance = 95%	Limite supérieure pour seuil de confiance = 95%
Constante	-2,05779368	3,07429533
Manipulateur 1	0,85612274	1,16367924

Dans ce tableau, nous pouvons d'abord trouver dans la colonne « coefficients » les valeurs de l'ordonnée à l'origine (qui correspond à la ligne « Constante ») et de la pente (qui correspond à la ligne « Manipulateur 1 »).

Dans la colonne « erreur-type », nous pouvons trouver les écart-types des paramètres a et b grâce auxquels il est possible de réaliser les tests de nullité de la pente et de l'ordonnée à l'origine. Les valeurs des deux statistiques de test sont données dans la colonne « statistique t ». La colonne « probabilité » donne la p-value pour chacun de ces deux tests, ce qui permet de conclure de la même manière que pour le test F :

- pour le test de l'ordonnée à l'origine, on a une p-value de $0,6537641 > \alpha = 5\%$. On est donc dans la zone de non rejet de H_0 ($b = 0$). On va donc pouvoir simplifier le modèle et faire passer la droite par 0.
- Pour le test de la pente, on a une p-value de $1,1094E-06 < \alpha = 5\%$. On est donc dans la zone de rejet de H_0 et d'acceptation de H_1 ($a \neq 0$). Y dépend de X à travers une relation linéaire de pente 1,00990099.

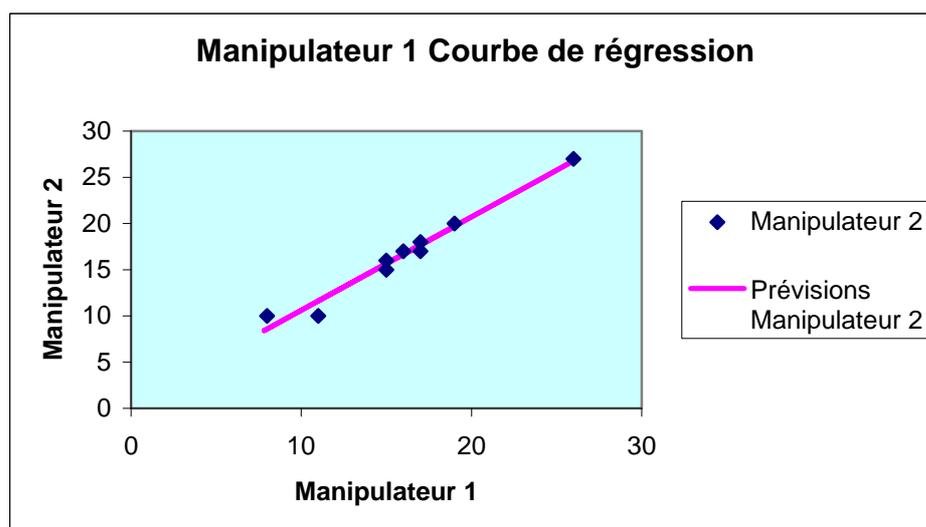
Comme pour le test F, on aurait pu donner la conclusion de ces deux tests en comparant la valeur calculée de la statistique t à la valeur critique qui se calcule en utilisant la fonction LOI.STUDENT.INVERSE appliquée aux arguments (α ; n-2).

La dernière partie du tableau donne les limites inférieures et supérieures des intervalles de confiance de a et b. Ces intervalles de confiance permettent aussi de comparer les paramètres à des valeurs : si la valeur à laquelle est comparé le paramètre se trouve dans les limites de l'intervalle alors on pourra dire que le paramètre et la valeur sont statistiquement identiques ; en revanche, si la valeur n'est pas dans les limites de l'intervalle, le paramètre ne sera pas du même ordre de grandeur que cette valeur. Par exemple pour l'ordonnée à l'origine, la valeur 0 se situe entre -2,05779368 et 3,07429533 ce qui nous permet de dire que l'ordonnée à l'origine est comparable à 0. Pour la pente, la valeur 0 se situe à l'extérieur de l'intervalle, la pente n'est donc pas comparable à 0.

Comme dans cet exemple nous nous demandons si les deux manipulateurs ont la même dextérité, nous voulons en fait savoir si la relation qui lie les deux séries de mesure est en fait $Y = X$, ce qui nous conduit donc – maintenant que nous avons vu que b est statistiquement égal à 0 – à regarder si la pente est statistiquement égale à 1. Nous allons donc comparer a avec la valeur 1. Pour cela soit on regarde si la valeur 1 se situe dans l'intervalle de confiance (ce qui est la cas), soit on remplace 0 par 1 dans la statistique de test ce qui donne $T = \frac{a - 1}{\sigma(a)}$. En utilisant les valeurs de a et de $\sigma(a)$ données par l'utilitaire d'analyse on obtient $t = 0,15224588$, qu'il faut comparer à la valeur critique de t qui vaut 2,36462256. Comme $0,15224588 < 2,36462256$, on ne rejette pas H_0 ($a = 1$).

Finalement, nous venons de vérifier que $Y = X$ ce qui signifie que les deux manipulateurs ont la même dextérité.

Associé à ce tableau, l'utilitaire d'analyse de la régression linéaire fournit aussi le graphique suivant dans lequel on trouve les points expérimentaux et la droite de régression calculée.

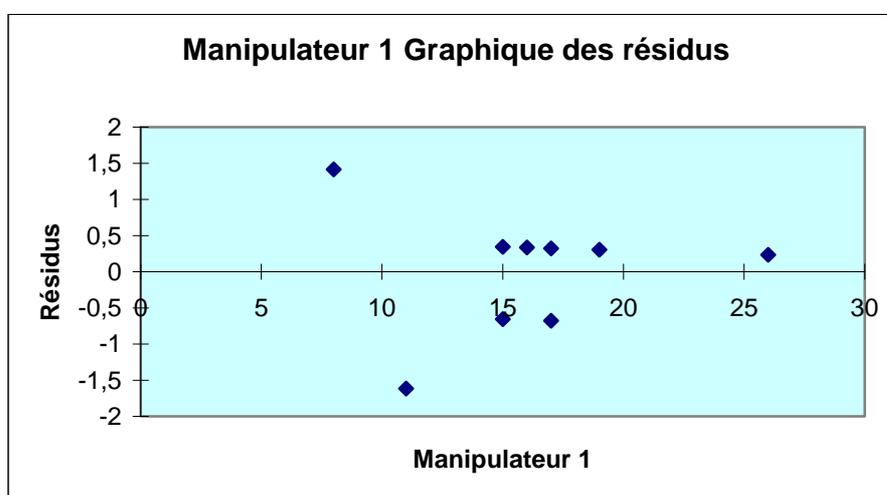


4) pour finir, le quatrième tableau donne les résidus, les résidus studentisés (ou normalisés) ainsi que les valeurs prédites de Y (Y'). Ce tableau

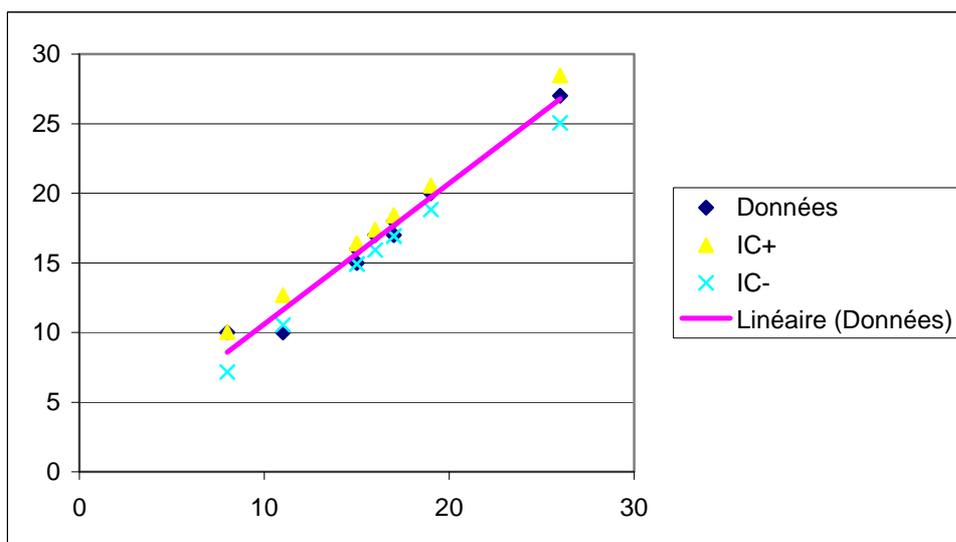
s'accompagne du graphique des résidus qui permet de faire tous les contrôles dont nous avons parlé dans le premier paragraphe.

ANALYSE DES RÉSIDUS

Observation	Prévisions Manipulateur 2	Résidus	Résidus normalisés
1	11,61716172	-1,617161716	-1,870426568
2	15,65676568	-0,656765677	-0,759622218
3	19,69636964	0,303630363	0,351182131
4	16,66666667	0,333333333	0,385536905
5	17,67656766	-0,676567657	-0,782525401
6	8,587458746	1,412541254	1,633760349
7	26,76567657	0,234323432	0,271020992
8	17,67656766	0,323432343	0,374085314
9	15,65676568	0,343234323	0,396988496



Pour finir, on peut alors déterminer les bornes positive (IC+) et négative (IC-) de l'intervalle de confiance de la droite.



II

1. Rappel des formules de calcul de la Régression

La droite de régression : $Y' = a + bX$

La constante : $a = \bar{Y} - b\bar{X}$

La pente : $b = r_{XY} \frac{S_Y}{S_X}$

Erreur type : $ET = \sqrt{\frac{\sum (Y_i - Y')^2}{n-2}} = \sqrt{\frac{SCR}{n-2}} = \sqrt{\frac{SCT}{n-1}} (1 - r^2 \text{ajusté})$

Le coefficient de *détermination* : $r^2 = \frac{SCE}{SCT}$

Et le coefficient de *détermination ajusté* : $r^2 \text{ajusté} = \frac{SCE - \frac{SCR}{n-2}}{SCT}$

D'où le calcul des erreurs-type en fonction des valeurs de X :

$$1) \text{ erreur-type de } b : \sigma_b = \frac{ET}{\sqrt{\sum (X_i - \bar{X})^2}} = \frac{ET}{\sqrt{SCT}}$$

$$2) \text{ erreur-type de } a : \sigma_a = ET \cdot \sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}} = ET \cdot \sqrt{\frac{\sum X_i^2}{n \cdot SCT}}$$

Rappel : $SCT = \text{Variance de } X \times n$.

2. Les Macro fonctions d'Excel

2.1 Données de base :

X	Y
80	6
36	2
134	15
99	6
65	4
28	1
11	1
58	4

2.2 Les calculs avec la Macro fonction « DROITEREG() »

b	0,1057653	-1,88075839	a
s(b)	0,01646563	1,22038848	s(a)
r ²	0,8730428	1,75085335	ET
F	41,2600207	6	ddl
SCE	126,482075	18,3929247	SCR
S(X ²)/n	5493,375	11306,875	SCT (X)

2.3 La Macro fonction « REGRESSION LINÉAIRE » en 3 parties :

2.3.1 La statistiques de la régression :

<i>Statistiques de la régression</i>	
Coefficient de détermination multiple	0,93436759
Coefficient de détermination R ²	0,8730428
Coeff. de détermination R ² ajusté	0,85188326
Erreur-type (ET)	1,75085335
Observations	8

2.3.2 Analyse de la variance :

	ddl	Somme des carrés	Moyenne des carrés	F	proba 0,05	f
SCE	1	126,4820753	126,4820753	41,260	0,0007	5,987
SCR	6	18,39292466	3,065487443			
SCT	7	144,875				

2.3.3 La statistique t vérifie si la pente est significativement différente de 0 :

	Coefficients	Erreur-type	Statistique t	Proba	Limite inf	Limite sup
Constante / a	-1,88075839	1,220388478	-1,5411	0,17422	-4,8669436	1,10542682
Variable X 1 / b	0,1057653	0,016465635	6,4234	0,0007	0,06547531	0,14605528

D'où l'équation de la droite de régression :

$$Y' = a + bX = -1,881 + 0,1058X$$

2.3.4 Contrôle des calculs en 4 parties :

2.3.4.1 Calculs de base sur X :

X	X ²
80	6400
36	1296
134	17956
99	9801
65	4225
28	784
11	121
58	3364
Moyennes =	5493,375
Variance =	1413,35938
SCT =	11306,875

2.3.4.2 Calcul des coefficients, des erreurs-type et de la statistique T :

	a	b
Coefficients	-1,88075839	0,1057653
ET		1,750853347
S(b)		0,016465635
S(a)	1,220388478	
T	-1,5411	6,4234

2.3.4.3 Calcul de t critique et des probabilités par les formules adéquates :

t : {Loi.student.inverse}. D'où la valeur du t critique = 2,447

$proba$: {Loi.student}. D'où la probabilité à 0,05% à ddl (1 ; 6) = 0,0007

NB : La probabilité de la constante $/a/$ – l'ordonnée à l'origine – n'a aucun sens. Il s'agit de vérifier si la pente de la droite de régression $/b/$ est différente de 0. En effet, si la pente était nulle, X et Y seraient confondues (superposées ou parallèles).

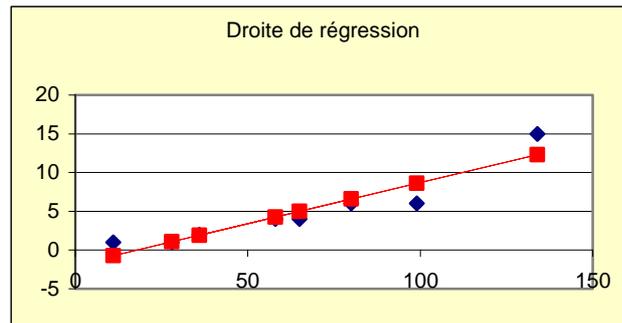
2.3.4.4 Détermination des seuils de signification :

$T > t = H_1$ significatif (Y différent de X = caractères propres)

$T < t = H_0$ non significatif (Y et X sont liés, proches ou apparentés)

Dans le cas présent, comme $T > t$, Y est différent de X.

D'où la droite de régression :



- la pente est ascendante
- et l'ordonnée à l'origine négative

Conclusion. Rien ne peut remplacer la pratique. L'exercice pratique est irremplaçable. On se forge une idée des techniques statistiques et on s'enrichit de toutes les connaissances du corpus traité.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

8

Formulaire

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce huitième chapitre, nous proposons un formulaire concernant les calculs des formules de base, de la décomposition de la variance, du calcul de la distance dt , des formes quadratiques, de l'inertie totale, des valeurs isotropes et des distances quadratiques, des corrélations et des ajustements, des résidus, des spectres de décomposition, et des différents que l'on peut être amenés à exécuter pour évaluer les différentes bases de données, comme les tests t , z ou F . Autant de calculs qui sont automatiquement exécutés par la Macro ou bien, que l'on va devoir programmer dans la feuille de calcul en fonction des besoins.

1. Formules de base

Primitives X et Y

Valeurs centrées x, y ($x = X - \bar{X}$ et $y = Y - \bar{Y}$)

Écart-type de X ($\sigma_x = \sqrt{\frac{\sum X^2}{N}}$) et écart-type de Y ($\sigma_y = \sqrt{\frac{\sum Y^2}{N}}$)

Écart centré réduit z : $z = \frac{x - \bar{x}}{\sigma_x}$ ou $z = \frac{x - \bar{x}}{\sqrt{npq}}$

Erreur-type de la moyenne : $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$ ou $S_{\bar{X}} = \frac{S_X}{\sqrt{n}}$

Test t d'un échantillon :
$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}}$$

Coefficients d'estimation $a = \frac{\sum xy}{\sum x^2}$ et $a' = \frac{\sum xy}{\sum y^2}$

Coefficient de corrélation $r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$ (cosinus)

Coefficient de détermination $r^2 = a.a'$

Coefficient de résistance $\rho = \sqrt{1 - r^2}$ (sinus)

Écart-type $\sigma_y = \sqrt{\frac{\sum y^2}{N}}$ de Y

Écart-type lié $u = \rho \cdot \sigma_Y$ (résidu quadratique moyen)

Droite de régression $Y' : Y' = ax + \bar{Y} \pm 2u$

Valeurs résiduelles de R : $R = \frac{Y - Y'}{u}$ (résidus quadratiques)

2. Décomposition de la variance

Variation totale : $\sum(Y - \bar{Y})^2 = \sum(Y - Y')^2 + \sum(Y' - \bar{Y})^2$

(La variation totale (σ_Y^2) = la variation résiduelle ($u^2 = \sigma_Y^2 (1 - r^2)$) + la variation expliquée ($r^2 \sigma_Y^2$).

Soit : $\sigma_Y^2 = (u^2 = \sigma_Y^2 (1 - r^2) + r^2 \sigma_Y^2)$

Variance totale = variance résiduelle + variance contrôlée

à savoir :

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y')^2 + \sum (Y' - \bar{Y})^2$$

D'où la racine de décomposition utilisée dans les calculs de la feuille d'estimation :

$$\boxed{(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})}$$

On pourra toujours s'assurer d'abord que pour chaque vecteur on a bien :

$$\boxed{(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})}$$

En vertu des théorèmes de Craig et de Cochran, des variantes du théorème de Pythagore, des relations de Thalès et des produits remarquables, on voit que ces variables de transformation sont indépendantes. Elles obéissent à une loi du χ^2 de R, avec un *ddl* (degrés de liberté) égal au nombre de lignes de la matrice d'estimation, avec, comme vecteur gaussien :

$$\boxed{(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})}$$

qui correspondent au carré d'une variable de loi LG (0,1).

Ce qui permet de vérifier la relation linéaire :

$$\boxed{\cos^2 + \sin^2 = 1 \Leftrightarrow r^2 + \rho^2 = 1}$$

exprimée par la formule :

$$\boxed{\sigma_y^2 = u^2 + c^2}$$

où $u = \rho \cdot \sigma_Y$ et $c = r \cdot \sigma_Y$

D'où l'indépendance des variables de transformation :

$$\boxed{\sum (Y - \bar{Y}) = 0} ; \boxed{\sum (Y - Y') = 0} ; \boxed{\sum (Y' - \bar{Y}) = 0}$$

et, par voie de conséquence, des variables de passage :

$$\boxed{\sum V_t = 0} ; \boxed{\sum V_r = 0} ; \boxed{\sum V_c = 0}$$

qui sont des variables centrées réduites :

1. $\boxed{V_t = \frac{Y - \bar{Y}}{\sigma_Y}}$ avec σ_Y (écart-type de Y)

2. $\boxed{V_r = \frac{Y - Y'}{u}}$ avec $u = \rho \cdot \sigma_Y$ (écart-type de résistance à la liaison ou résidu quadratique moyen)

$$3. \boxed{Vc = \frac{Y' - \bar{Y}}{c}} \text{ avec } c = r \cdot \sigma_Y \text{ (écart-type de liaison contrôlée)}$$

$$\text{de forme classique } z : \boxed{z = \frac{Y - \bar{Y}}{\sigma_Y}}$$

Le spectre de la matrice de passage R qui obéit à une loi du χ^2 de R définie par la formule de calcul de la distance quadratique dt de Y en X (« opération de polarisation » par transformation de la forme bilinéaire) :

$$\boxed{dt = \sqrt{\frac{1}{2} [(Vc^2 + Vr^2) - Vt^2]}}$$

Ce spectre rend parfaitement compte de la densité et de la qualité de la résistance ou de la liaison des éléments constitutifs des vecteurs linéaires portés par la matrice d'estimation de Y en X, quelle que soit la provenance des items, qu'ils soient tirés de la TDF ou du Lexique (les Tables de contingence), ou encore qu'ils soient considérés globalement ou individuellement.

La somme des carrés de dt est toujours égale à $n/2$:

$$\boxed{\sum (dt)^2 = \frac{n}{2}}$$

n étant le nombre de lignes ou d'items factoriels.

Il s'ensuit que dt joue le même rôle que le carré de la norme pour le produit scalaire. D'où la polarisation qui en résulte en corollaire, avec la formation des vecteurs isotropes. Toute forme bilinéaire symétrique est équivalente à la donnée d'une forme quadratique.

Nous verrons plus loin, au § 7, 6, que l'inertie totale de Vr^2 de Y/X correspond à l'inertie totale de dt^2 de X/Y ; et que $Vr^2/2$ de l'un est égal à dt^2 de l'autre, et inversement, avec la plus stricte correspondance des valeurs linéaires ou factorielles.

NB. Il suffit de diviser le radical de dt par 0,5 pour que la somme des $dt^2 = n$. Cela ne changerait rien aux valeurs supérieures à 1. En revanche, le seuil des valeurs immédiatement inférieures à 1 serait élargi d'un angle de 30°. Mais la limite référentielle de 0,500 est toujours valable.

3. Le pronostic

- X, c'est la variable explicative dite *prédicteur* : elle est déterminante ou référentielle
- Y, c'est la variable expliquée dite *critère* : elle est déterminée ou estimative
- Y', c'est la droite d'estimation (d'ajustement ou de régression) : elle est estimée.

4. Le calcul de la distance quadratique dt

Quelques précisions sur la formule de calcul de la distance quadratique dt (que le lecteur pourra compléter en se reportant à des ouvrages de mathématiques, d'analyse linéaire, de calcul vectoriel ou matriciel, de mécanique ou encore à des ouvrages de statistique).

La formule que nous proposons pour le calcul de la distance dt , équivalente au calcul du χ^2 de R, mérite explication et justification.

Compte tenu du principe d'additivité de la variance,

Vu le théorème de Pythagore appliqué à la trigonométrie et à la projection orthogonale,

Vu la transformée de Fourier,

Vu les théorèmes de Cauchy-Schwarz, de Craig et de Cochran,

En vertu des produits remarquables et de la symétrie du produit scalaire,

La formule du calcul de la distance quadratique dt est immédiatement justifiée :

$$dt = \sqrt{\frac{1}{2} [(Vc^2 + Vr^2) - Vt^2]}$$

En vertu du principe d'additivité de la variance, et vu le passage des valeurs primitives X et Y (afférentes à la matrice d'estimation de Y en X) aux formes quadratiques de la matrice de transformation et aux valeurs centrées réduites de la matrice de passage R, on peut résumer ainsi la transformation logique qui aboutit à la formule de la distance quadratique dt de Y en X :

:

1) principe d'additivité de variation : $Vt^2 = Vr^2 + Vc^2$

2) métrique trigonométrique : $\cos^2 + \sin^2 = 1$

3) transformation des variables : $\sum (Y - \bar{Y}) = 0$; $\sum (Y - Y') = 0$; $\sum (Y' - \bar{Y}) = 0$

4) indépendance des variables : $\sum Vt = 0$; $\sum Vr = 0$; $\sum Vc = 0$

5) spectre des valeurs linéaires : Vt , Vr et Vc (somme et moyenne = 0 et l'écart type = 1)

6) d'où les vecteurs et les cônes isotropes d'une forme quadratique en dimension infinie.

Le calcul de la distance quadratique dt suit la métrique euclidienne usuelle : elle est normalement définie dans l'intervalle ($0 \leq dt \leq 1$) (suivant le principe de la loi du χ^2) et devient hautement significative (positive ou négative suivant la valeur algébrique) au-delà de la limite ($dt \geq 1$).

Rappel de la formule de l'inégalité de Cauchy-Schwarz qui permet de fixer la norme dans l'intervalle de variation :

Soit E un espace vectoriel réel muni d'un produit scalaire.

Puisque $\langle x, x \rangle \geq 0$ pour tout $x \in E$, on peut en considérer la racine carrée :

$$\|x\| := \sqrt{\langle x, x \rangle}$$

D'où l'inégalité triangulaire :

$$\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in E$$

En vertu de l'inégalité de Cauchy-Schwarz, si le déterminant est ≤ 0 , on a :

$$\langle x, y \rangle^2 \leq \|x\|^2 \|y\|^2$$

l'égalité n'ayant lieu que si x et y sont liés.

D'où l'inégalité de Cauchy-Schwarz :

$$\langle x, y \rangle^2 - \|x\|^2 \|y\|^2 \leq 0$$

Suivant l'inégalité de Cauchy-Schwarz, si 2 vecteurs x et y ne sont pas nuls, on a :

$$\frac{|\langle x, y \rangle|}{\|x\| \|y\|} \leq 1$$

Il existe par conséquent un et un seul $\theta \in [0, \pi]$ tel que :

$$-1 \leq \cos \theta = \frac{\langle x, y \rangle}{\|x\| \|y\|} \leq 1$$

Or, l'angle θ entre les vecteurs x et y est dit « non orienté ».

D'où la relation qui exprime le produit scalaire en fonction de la norme :

$$\langle x, y \rangle = \frac{1}{2} (\|x + y\|^2 - \|x\|^2 - \|y\|^2)$$

La démonstration est immédiate :

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \|x\|^2 + \|y\|^2 + \langle x, y \rangle + \langle y, x \rangle = \|x\|^2 + \|y\|^2 + 2 \langle x, y \rangle$$

Mais, en vertu du produit remarquable (\Rightarrow démonstration du théorème de Cauchy) :

$$(a + b)^2 > a^2 + b^2$$

étant donné que :

$$\boxed{(a + b)^2 = a^2 + b^2 + 2ab}$$

D'où la proposition de la formule de dt :

$$\boxed{dt = \sqrt{\frac{1}{2} [(Vc^2 + Vr^2) - Vt^2]}}$$

Et donc :

- 1) Si $0 \leq dt \leq 1$, la liaison entre X et Y est normale, la distance euclidienne s'inscrit normalement dans le cercle $(0, 1)$.
- 2) Si $dt \geq 1$, Y est de moins en moins lié à X , ou de moins en moins expliqué par X , de sorte que le point d'ancrage de dt est extérieur au quartier positif du cercle $(0, 1)$. Cas des valeurs « aberrantes », à forte déviance.
- 3) Le cas de $dt < 0$ n'est pas envisageable, pour des raisons évidentes de calcul de distances euclidiennes.
- 4) Le cas où $dt = 0$ n'est guère envisageable non plus pour la bonne raison que cela signifierait que $X = Y$, et donc que la distance quadratique dt est réduite à 0 par la superposition des deux droites. Dans ce cas, la droite d'estimation Y' n'existe pas. Il est aisé d'en faire la vérification à l'aide de la MACRO, d'autant plus que l'on peut facilement inverser les valeurs de X et de Y .

5. Formes bilinéaires et formes quadratiques

L'intérêt des *formes bilinéaires* de type $s(x,y)$ est de les situer dans la théorie de la Relativité restreinte où le produit scalaire est remplacé par la forme bilinéaire, de telle sorte que la transformation en *forme quadratique* $q(x) = s(x,y)$, par le biais du polynôme homogène de degré 2, suivant la « *règle de dédoublement des carrés* », dite aussi « *opération de polarisation* » :

$$\boxed{s(x,y) = \frac{1}{2} [q(x+y) - q(x) - q(y)]}$$

NB : s est dite *forme polaire* de q , étant donné que $q(x)$ joue le même rôle que le carré de la norme pour le produit scalaire.

En vertu de la linéarité de l'intégrale, s est une forme bilinéaire symétrique :

$$\boxed{q(P) = \int_0^1 P(x)^2 dx}$$

En effet, on a :

$$\boxed{s(P, Q) = \frac{1}{2} [q(P + Q) - q(P) - q(Q)] = \int_0^1 P(x)Q(x)dx}$$

On aboutit alors à la formation d'un double cône isotrope inversé au centre de gravité $O(\bar{x}, \bar{y})$, où s'exprime l'orthogonalité vectorielle de décomposition de la variance (suivant le théorème de Pythagore), et de réduction en vecteurs de densité (V_t, V_r, V_c).

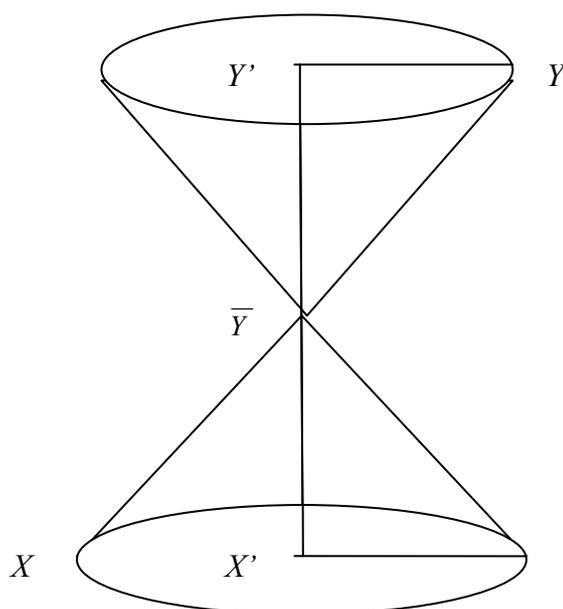
D'où la considération de la parabole symétrique définie par $y = x^2$ (qui contrebalance largement la courbe en forme de cloche de Laplace-Gauss).

La matrice quadratique q se réduit à un espace vectoriel de rang n ($\text{rg} = n$, nombre de facteurs ou de lignes de la variable) et d'inertie totale $(n, 0)$, reflétant l'indépendance des variables :

$$\boxed{\sum V_t = 0} ; \boxed{\sum V_r = 0} : \boxed{\sum V_c = 0}$$

D'où la valeur des spectres de transformation des régressions et des représentations spectrales des distances quadratiques (ou des moindres carrés) au centre de gravité du cercle (représentés dans un cercle sous la forme d'un « carré de côté 1 »), permettant de visualiser les distances quadratiques des résidus (*voir infra*).

Les points représentant Y se situent sur la surface d'une double cône inversé au centre où figure la valeur moyenne de \bar{Y} et la valeur Y' de l'estimation se projetant sur l'axe central. Il s'ensuit que les valeurs positives des résidus de Y sont sur le cône supérieur et les valeurs négatives (se rapportant à X) sont sur le cône inférieur.



Pour chaque vecteur, le triangle rectangle, ayant pour angle droit les côtés YY' et $Y'\bar{Y}$, est produit par les valeurs de décomposition de la variance. D'où le cône isotrope sur lequel se projettent toutes les valeurs de Y faisant suite à la décomposition de la variance dans l'analyse de la régression.

NB : Dans tout système quadratique d'estimation de la régression $- [s(x,y) = q] -$, on remarque que les valeurs relatives au *critère* Y se projettent sur le cône supérieur, alors que les valeurs relatives au *prédicteur* X se projettent sur le cône inférieur inversé. De telle sorte que la

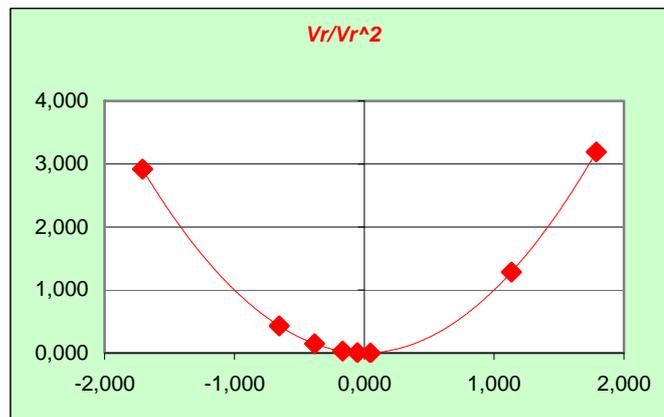
totalité des vecteurs isotropes forme une courbe hélicoïdale qui s'enroule comme une spirale autour de l'axe $Y'\bar{Y}$, axe central des deux cônes. Lorsque $X = Y$, le cône se réduit à une simple droite : il n'y a pas de résidus.

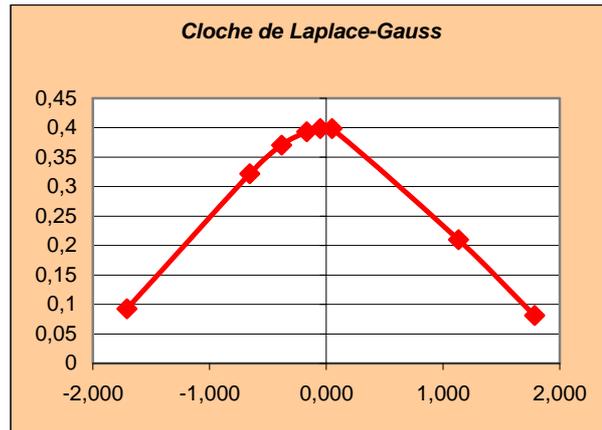
On aboutit ainsi au classement rigoureux et « robuste » de tous les éléments factoriels qui se projettent au gré des valeurs dans toutes les zones de découpage du cône, faisant alors la part belle « aux valeurs qualitatives » de la distribution.

Remarque : On appelle **cône isotrope** l'ensemble $I(q)$ des vecteurs défini par $I(q) := \{x \in E \mid q(x) = 0\}$. Tel est présentement le cas, puisque les sommes de V_t , V_r et V_c sont nulles ; que les sommes des carrés sont égales à n , le nombre de lignes ou de facteurs ; que la somme des carrés de dt est aussi égale à n , le nombre de lignes ou de facteurs. En fonction des calculs de décomposition de la variance, la forme bilinéaire $s(x,y)$ se projette sur une base orthogonale.

6. Inertie totale, vecteurs isotropes et distances quadratiques

- 1) L'analyse des résidus est fondamentale dans l'estimation des données.
- 2) La distance dt et les spectres des vecteurs isotropes
- 3) Les graphes d'inertie des distances entre les résidus et la variation totale et la variation contrôlée
- 4) Les carrés d'inertie et les graphiques isotropes et paraboliques : $(0,n)$ pour V_r/V_r^2 et $(x,n/2)$ pour dt/dt^2



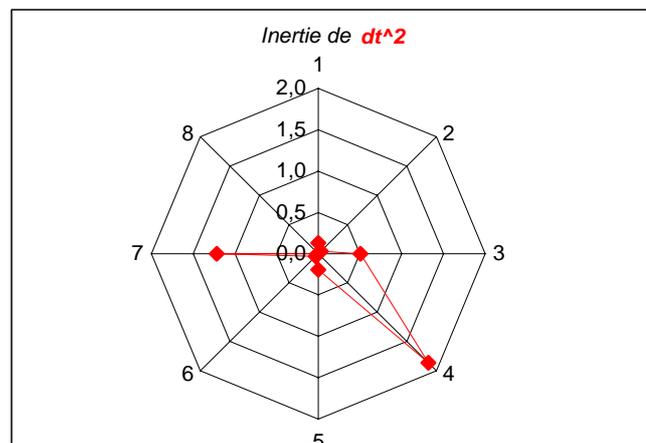
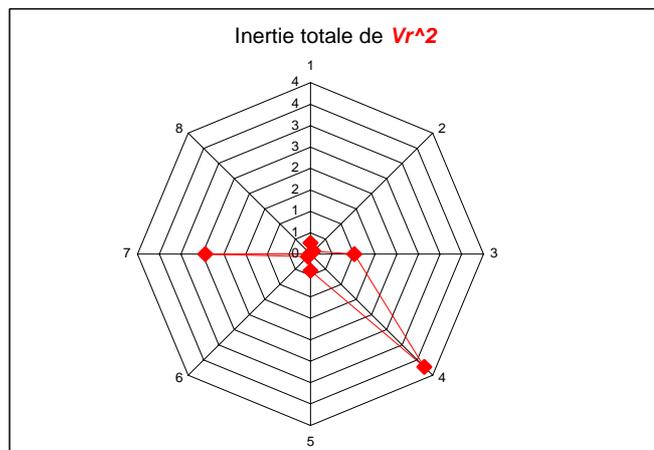


5) Parallélisme inversé des graphiques isotropes entre Vr^2 de X en Y et dt^2 de Y en X .

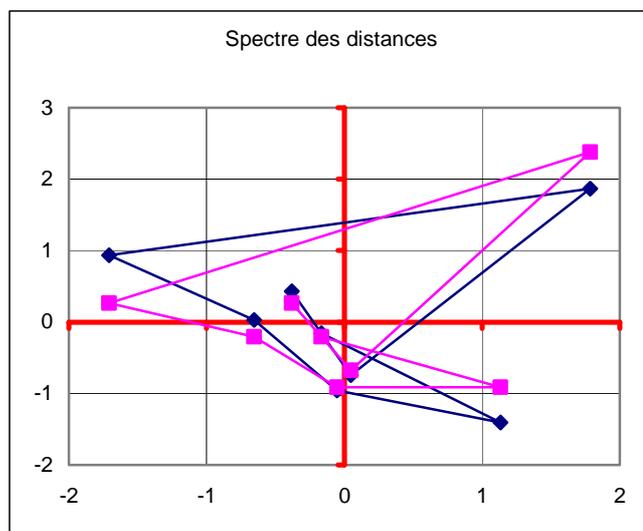
6) dt^2 de Y en $X \Leftrightarrow (Vr^2)/2$ de X en Y .

NB : Les calculs de dt de Vr sont immédiatement liés. D'où l'importance du calcul des résidus, qui ne sont que les « résistances quadratiques » aux liaisons estimées des deux variables.

D'où l'intérêt du calcul des distances quadratiques linéaires de X en Y et de Y en X , ou de X/Y et de Y/X , our pouvoir observer la distribution des vecteurs isotropes :



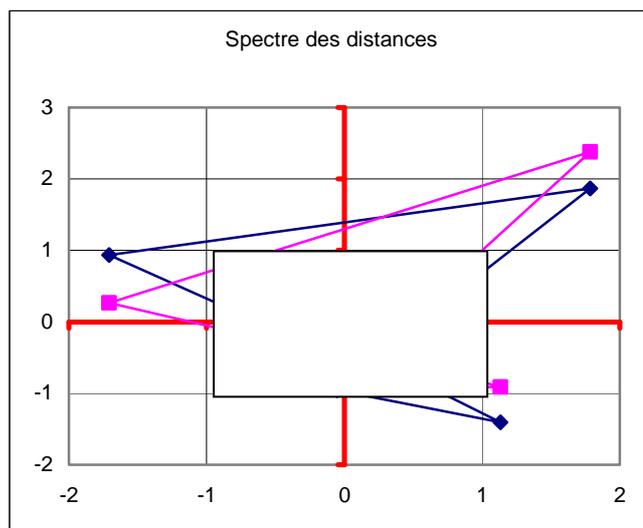
7) profils d'inertie des spectres de décomposition et distances quadratiques des résidus.



Le spectre des distances quadratiques montre combien les trois vecteurs de décomposition de la variance sont liés : la distance de Y par rapport à X ou de X par rapport à Y (bleu pour X et rouge pour Y).

Et, par voie de conséquence, quelle est l'importance des valeurs des distances dt , qui projettent la spirale autour des cônes isotropes.

Le quadrillage montre les 3 éléments qui se projettent à l'extérieur du carré de normalité, de centre 0 et de rayon 1. Pour des raisons techniques, il n'est pas possible de délimiter ce carré de côté 1. Dans le meilleur des cas, on peut l'occulter :



Ainsi les trois facteurs sont immédiatement identifiés.

Bien entendu, la discrimination des éléments factoriels, et notamment ici des 3 éléments remarquables, est faite par le rangement par ordre décroissant des valeurs de dt pour les

valeurs positives de Vr , et inversement, par ordre croissant de valeurs de dt pour les valeurs négatives de Vr . (Voir, par exemple, l'étude du *Petit Chaperon rouge* de Perrault).

7. Formules de calcul des corrélations et des ajustements :

Y' de Y / X	X' de X / Y
r	r
r^2	r^2
$\rho = \sqrt{1 - r^2}$	$\rho = \sqrt{1 - r^2}$
σ_Y	σ_X
$u = \rho \cdot \sigma_Y$	$u' = \rho \cdot \sigma_X$

– $\boxed{\cos^2 + \sin^2 = 1}$; $\boxed{r^2 + \rho^2 = 1}$

– $\boxed{x = \frac{r+1}{2}}$ (distance euclidienne, par excellence)

– $\boxed{y = \frac{\rho+1}{2}}$ (distance euclidienne, par excellence)

8. Calcul des résidus

Calcul des écarts centrés réduits

Une fois la feuille préparée, le calcul des écarts réduits se fait automatiquement. On écrit ou on recopie la formule appropriée dans la cellule D4 :

$$\boxed{=(C4-B4*\$D\$1)/RACINE(B4*\$D\$1*\$D\$2)}$$

que l'on étend ensuite à toute la colonne.

9. Spectre de décomposition

Le calcul des densités des vecteurs centrés réduits Vt , Vr et Vc :

1. $\boxed{Vt = \frac{Y - \bar{Y}}{\sigma_Y}}$ avec $\boxed{\sigma_Y}$ (écart-type de Y)

2. $Vr = \frac{Y - Y'}{u}$ avec $u = \rho \cdot \sigma_Y$ (écart-type résiduel ou résidu quadratique moyen)
3. $Vc = \frac{Y' - \bar{Y}}{c}$ avec $c = r \cdot \sigma_Y$ (écart-type de liaison contrôlée)

Si les variables du spectre sont indépendantes, elles doivent être nulles (comme les primitives de variance).

On pourra s'assurer que :

$$1. \sum Vr = 0 \quad \text{et} \quad \sum (Vr)^2 = n$$

$$2. \sum Vc = 0 \quad \text{et} \quad \sum (Vc)^2 = n$$

$$3. \sum dt = 0 \quad \text{et} \quad \sum (dt)^2 = n$$

d'où la valeur de la distance quadratique dt (équivalente en somme d'un χ^2) :

$$4. \sum (dt)^2 = \frac{n}{2}$$

d'où les valeurs référentielles du paramètre dt : $dt = 1$ et $dt = 0,500$.

Alors les vecteurs du spectre (suivant la transformée de Fourier, et conformément au théorème de Craig, qui n'est autre qu'une variante du théorème de Pythagore) – qui ne sont rien moins que les formes quadratiques de la matrice d'estimation –, répondent à la généralisation de l'additivité du χ^2 .

NB : Les coefficients des termes perpendiculaires sont divisés par 2, ce qui restreint le champ des éléments hautement significatifs (les éléments thématiques prépondérants).

D'où l'*homoscédasticité* ou « *distribution anarchique du nuage de points de Vr* » pour un écart-type σ conditionnel et constant (elle est fondamentale dans l'étude des résidus). Les graphes des résidus en fonction des variables explicatives ne doivent laisser apparaître aucune tendance.

NB : Comme nous avons affaire à des « résultats constants », nous avons aussi affaire à des *lois de mécanique*, amplement décrites en *Algèbre linéaire*.

En résumé : on vérifiera que la somme des valeurs des colonnes D8 à I8 est nulle.

Rappel : le calcul de l'indice de distance quadratique dt de chaque vecteur est donné par la formule suivante :

$$dt = \sqrt{\frac{1}{2} [(Vc^2 + Vr^2) - Vt^2]}$$

Cette formule n'est rien moins que le calcul du χ^2 de Fisher propre à chaque vecteur.

NB : Lors de l'étude du vocabulaire d'une des variables considérée sous l'angle de la régression (simple ou multiple), on placera, après les valeurs centrées réduites de z , les valeurs centrées réduites de décomposition de la variance, V_t , V_r et V_c , de telle sorte que l'on peut détacher les différentes strates du vocabulaire correspondant aux impératifs d'écriture et de composition :

- a) dans le cône supérieur, où les valeurs de dt sont rangées par ordre décroissant pour un $V_r \geq 0$, la spirale découpe le vocabulaire de prédilection en 3 zones :
 - en zone à dominante thématique pour un $dt \geq 1$,
 - en zone de sous-dominante thématique pour $0,5 \leq dt \leq 1$,
 - et en zone de remplissage (par nécessité d'enrichissement thématique, descriptif, rationnel ou autres fioritures et modulations), lorsque $0 \leq dt \leq 0,5$.
- b) dans le cône inférieur inversé, où les valeurs de dt sont rangées par ordre croissant pour un $V_r \leq 0$, la spirale découpe le vocabulaire en zones symétriques à celles de la partie supérieure. Néanmoins, dans ces zones à valeurs « négatives » (en étroit rapport avec le critère X), il faut remarquer les éléments à forte densité absolue ($z \geq 1$) : il s'agit généralement de facteurs qui figurent dans les basses fréquences du vocabulaire, mais qui jouent un rôle déterminant du point de vue esthétique, rationnel et discursif. Ce sont, par exemple, des vocables de fréquence 1, souvent des hapax, qui entrent dans la richesse descriptive du texte.

Les deux parties de la spirale, qui se projettent sur les deux cônes inversés, forment la parabole intégrale des résidus donnée par V_r / V_r^2 : la branche positive correspond à la partie qui se projette sur le cône supérieur et la branche négative, à la partie qui se projette sur le cône inférieur. Aussi les valeurs positives de la régression préfigurent-elles l'ancrage de Y dans la partie supérieure du cône isotrope et les valeurs négatives, l'ancrage de X dans la partie inférieure (du cône inversé).

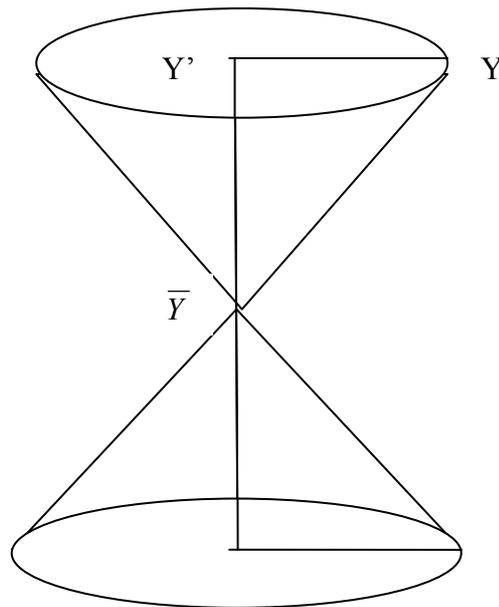
Les deux cônes sont portés par l'axe $Y\bar{Y}$, qui se prolonge en $Y'X'$, compte tenu des valeurs négatives de V_r qui renvoient directement l'estimation aux valeurs du « prédicteur X ».

Bien qu'opposés par le sommet, les deux cônes sont complémentaires, ils embrassent la totalité des composantes factorielles prises en charge par la *Régression* (simple ou multiple). Si le cône supérieur ne retient que les valeurs positives qui définissent l'identité de la variable Y , le cône inférieur, lui, ne retient que les valeurs négatives qui font la différence avec le prédicteur X . Or, pour reprendre Aristote qui disait que toute question est une « *question d'identité ou de différence* » (Voir *Les Topiques*, I, 6), la définition du critère Y se fait en fonction du prédicteur X , par corrélation, de sorte que le tout éclaire la partie et que la partie se reflète dans le tout. La définition englobe le contenu des deux cônes, où se dessinent tout à la fois « l'identité et la différence », l'une n'allant jamais sans l'autre. La question étant globale et unitaire, elle est à la base des principes rationnels : *le principe d'identité, le principe de contradiction et le principe du tiers exclu*.

Il n'y a pas de ligne de partage des eaux. C'est l'ensemble des valeurs d'estimation qui constitue le fil d'Ariane, fil de trame et fil du discours. L'identité fait la différence et la différence fait l'identité. C'est le principe même de l'analyse ($\alpha\nu\alpha\lambda\upsilon\epsilon\iota\nu$) que de séparer pour comprendre les mécanismes d'unité, de fonction et de fonctionnement. La montre, ce n'est pas un ensemble de pièces dans un chapeau, mais les pièces assemblées en tout, fonctionnant en cadence suivant le principe créateur, dans un but bien précis, qui n'est pas de « donner

l'heure », comme on le prétend, mais de permettre à son utilisateur de fixer le centre trigonométrique de sa vie, et donc de se situer et de s'orienter dans l'espace et dans le temps.

Aussi toutes les valeurs de z , de Vr (et aussi de Vt ou de Vc), sont-elles réorganisées suivant l'ordre décroissant ou croissant des valeurs de dt pour décrire la spirale intégrale des valeurs de Y s'enroulant autour des cônes isotropes. Cette spirale est centrée sur un point de rupture au point 0 (zéro), à la pointe d'inversion des deux cônes, où l'inertie de la matrice est totale, les distances quadratiques nulles et la dynamique de régression entre « *prédicteur X* » et « *critère Y* » inversée.



D'où l'importance du calcul des distances quadratiques de dt qui permettent de découper le vocabulaire de chaque variable en zones et en strates reflétant les choix des vocables en vertu des principes rationnels et en fonction des qualités artistiques et esthétiques inhérentes au texte et au discours.

La meilleure façon de découvrir les secrets de l'art et de la technique, c'est encore de s'adonner à l'étude de la régression (simple ou multiple) des variables d'un corpus, pour découvrir qu'au-delà de l'harmonie globale d'un corpus, chaque variable est unique, indépendante et autonome.

On pourra consulter ou s'inspirer des études des contes de Perrault, comme le *Petit Chaperon rouge*, ou des contes de Miguel Torga, comme *Vindima*, *Cavaquinho*, *A Paga*, ou autres, pour mieux assimiler la technique d'analyse de la régression, qui n'est ni simple ni évident de prime abord, mais qui est parfaitement abordable et toujours utile, voire indispensable. Disons en outre qu'elle mérite bien le titre de « scientifique » (*stricto sensu*), étant le poids des structures mathématiques qui lient les trois vecteurs de décomposition de la variance de Y en fonction de X .

Le calcul des distances quadratiques dt , suivant la règle du dédoublement des carrés, correspond à « *l'opération de polarisation* » du système bilinéaire, lequel est parfaitement défini et déterminé par la connaissance des valeurs quadratiques de dt reflétant la distribution des valeurs de Y sur le cône isotrope.

[Voir, *supra*, 6,7, les profils d'inertie des spectres de décomposition : distances quadratiques des résidus.]

Pour de plus amples explications, le lecteur avisé peut se reporter aux chapitres de *mécanique* et *d'algèbre linéaire* qui traitent de la « *Relativité restreinte* ». Il va sans dire que la mécanique et l'algèbre linéaire occupent une place prépondérante dans toutes les branches scientifiques (physique, chimie, économie, statistique, informatique...): elles sont « *incontournables* » dans la mesure où l'algèbre et la géométrie se mêlent constamment pour solliciter en permanence l'imagination. Les graphes permettent de repérer immédiatement les items affectés d'une « *valeur remarquable* », sachant qu'un fort résidu peut indiquer une « *valeur aberrante* », mais qu'une valeur peut être « *aberrante* » sans que son résidu soit important. C'est le cas notamment des vocables de faible fréquence alors qu'ils concernent des populations de fortes densités, comme les hapax ou les vocables de fréquence 1 ou 2, ou autres (voir les *Tables de contingence*, comme les *Lexiques* et les *TDF*): ces vocables, ayant une densité z hautement significative, peuvent être relégués, pour des raisons techniques, dans la zone négative du cône inversé, alors que ce sont des vocables qui jouent un rôle de premier plan du point de vue esthétique et discursif.

NB : Il y a une différence fondamentale entre la Régression simple et la Régression multiple. Dans la Régression simple, les variables sont appariées, et donc réversibles. Dans la Régression multiple, chaque variable est comparée à l'ensemble des autres variables réunies, et donc irréversibles. Pourquoi ? Dans le cas de la Régression multiple, notamment au niveau des basses fréquences, le vocabulaire du « prédicteur X » est dispersé ou éparpillé sur toute la Table de contingence. Il est donc nécessaire de recourir aux densités absolues de z pour étudier exhaustivement les valeurs du « critère Y ». Prenons l'exemple des 8 contes en prose de Perrault : le vocabulaire de chaque variable est estimé par rapport au « vocabulaire correspondant » de l'ensemble des variables du corpus. D'où l'importance de considérer les densités de z pour les populations à forte densité dans les basses fréquences, qui sont reléguées en fonction des résidus « négatifs » ($V_r \leq 0$) dans les « zones atypiques » ou différentielles du cône inférieur.

La lecture des données analytiques doit être globale. On voit alors que le nombre, que l'on croyait (naïvement sans doute) purement « *quantitatif* », souligne en permanence « *la valeur qualitative* » de toutes les composantes factorielles qui forment la masse lexicale et texture du discours.

S'il fallait résumer la fonction de l'analyse factorielle discriminante, l'AFD, il faudrait, sans aucune hésitation, dire qu'elle focalise toute l'attention sur l'hypertexte au moyen d'un éclairage exceptionnellement réaliste et objectif pour dévoiler les contours de toutes les composantes inhérentes au lexique, au texte et au discours.

La meilleure façon de s'en convaincre, c'est, de toute évidence, (et Dieu sait si *l'évidence* est fondamentale en philosophie et en logique), de s'adonner à la pratique du calcul de la *Régression* pour fixer les valeurs factorielles et dépasser le cadre purement quantitatif (qui hante les esprits obnubilés).

10. TEST *t* des MOYENNES

Test *t* de STUDENT-FISHER

appliqué à la comparaison de variables appariées

$$T = \frac{D}{\sigma_D} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{VarX + VarY}{N(\text{couples})}}}$$

t = loi.student.inverse (α ; *ddl*) avec $\alpha = 0,005$ ou $0,01$ et *ddl* = $2(N - 1)$.

11. Analyse de la variance interclasse et intraclasse appliquée à la comparaison simultanée de plusieurs moyennes ou à la décomposition de la variance

La décomposition de l'échantillon sui la loi de décomposition de la régression :

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y')^2 + \sum (Y' - \bar{Y})^2$$

donne les *ddl* suivants :

$$\text{Var. Totale } (n - 1) = \text{Var Résiduelle } (k - 1) + \text{Var. Estimée } (n - k)$$

Ou, pour une matrice multiple :

$$\sigma^2 = \rho^2 \sigma^2 + r^2 \sigma^2$$

$$\sigma^2 = \mathbf{u}^2 + \mathbf{c}^2$$

d'où les *ddl*

$$(n - 1) = (k - 1) + (n - k)$$

d'où la formule de calcul du F de Snédécour (en utilisant les Macro) :

$$F = \frac{\frac{r^2}{k-1}}{\frac{\rho^2}{n-k}}$$

$f = \text{inverse.loi.F}(\alpha ; k-1 ; n-k)$

avec k = nombre de colonnes pour l'interclasse ou intergroupe au numérateur et n = nombre de lignes ou de paires pour l'intraclasse au dénominateur.

Exemple : Analyse de la variance de la Régression Y' de Y en X (8 lignes) avec $k-1 = 2 - 1 = 1$ ddl pour le numérateur et $n-k = 8 - 2 = 6$ ddl pour le dénominateur:

18,109	2,299	15,810	
	$u^2 / n-k$	$c^2 / 1$	
	0,383	15,810	F = 41,260
r^2	$\rho^2 / (n-k)$		
0,873	0,021		F = 41,260
ddl	p value	f	
(0,05%;1 ;6)	0,05	5,987	<i>H₁ = significatif de caractères propres</i>
(0,01%;1 ;6)	0,01	13,745	
(0,0007%;1 ;6)	0,00067	41,313	

Analyse de la variance à l'aide des *Utilitaires* appliquée au même corpus et calculée selon les règles :

1) calculs sur la feuille de Macro

T^2	R^2	C^2
1,266	0,337	2,909
8,266	0,005	8,692
102,516	7,334	55,009
1,266	6,708	13,801
0,766	0,988	0,014
15,016	0,007	14,397
15,016	2,949	31,274
0,766	0,064	0,386
144,875	18,393	126,482
144,875	3,065	F
		41,260
k - 1 =	n - k =	
1	6	
r2	$\rho^2 / 6$	
0,873	0,021	41,260
0,05	f =>	5,987
0,01		13,745
0,00067	<= p value	41,313

2) calculs par la fonction « *régression* » des Utilitaires :

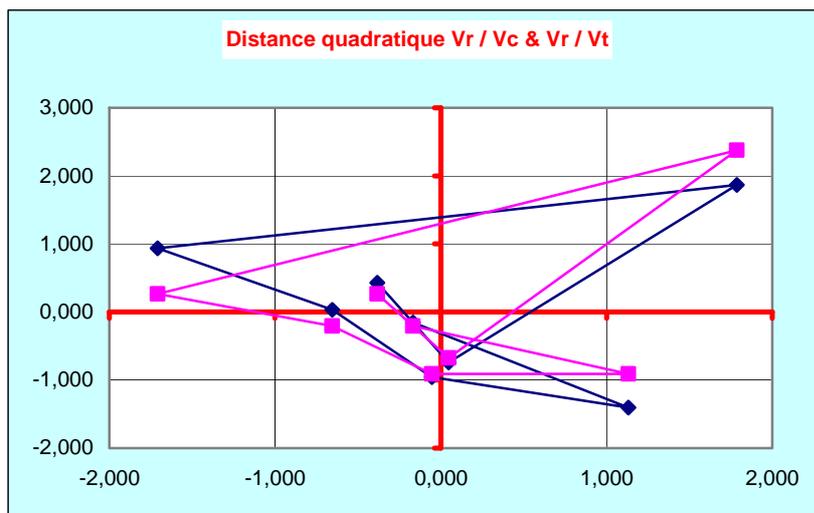
ANALYSE DE VARIANCE						
	Degré de liberté	Somme des carrés	Moyenne des carrés	F	P value	f
Régression	1	126,482	126,482	41,260	0,05	5,987
Résidus	6	18,393	3,065		0,01	13,745
Total	7	144,875			0,00067	41,313

avec : =Inverse.loi.F(0,00067;1 ;6) on obtient : 41,313

Test hautement significatif des distorsions qui affectent les 2 distributions. La décomposition de la variance montre quelle est l'importance des valeurs des résidus qui se glissent entre les valeurs de la variance de Y et de Y' :

Y	Y'	(Y - Ymoy)	(Y - Y')	(Y' - Ymoy)
18,109	15,810	18,109	2,299	15,810
			0,383	

D'où le schéma des *distances quadratiques des résidus* mettant en évidence les éléments aberrants (Voir *méthode des moindres carrés*) :



La Macro donne 3 éléments aberrants : 2 liés à Y (D3 et D7) et 1 à X (D4), et 5 éléments centrés. D'où les valeurs des distances $dt \geq 0,500$.

12. TEST de comparaison des *mêmes* VARIANCES

Test F de Fisher-Snédecor

Comparaison de 2 « populations » ayant la même variance

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

$f = \text{inverse.loi.F}(\alpha; n_1 - 1; n_2 - 1)$.

13. Analyse de la variance pour comparaison simultanée

Exemple : Les prix d'un même produit dans 4 (k) villes, pour 8 (n) observations, soit 32 (N) relevés.

X1	X2	X3	X4
9	10	10	9
8	11	13	8
10	9	12	10
9	10	11	7
8	12	11	9
10	11	10	10
11	10	12	9
11	11	11	8

k = colonnes = 4 n = lignes = 8 N = total = 32

ddl numérateur : k - 1	4 - 1 = 3
ddl dénominateur : N - k	32 - 4 = 28

Moyenne
Variance

9,5	10,5	11,25	8,75	9,667	interclasse
1,429	0,857	1,071	1,071	1,107	intraclasse
				F =	8,731
				f =	2,947
				f =	8,735
				(0,0003)	0,0003

VAR / k	1,43	0,86	1,07	1,07	4,43		
VAR / k		k = 4			1,107	intraclasse	intergroupe
VAR / TOTAL (N)		N = 32			1,208		N = total
VAR pour n		n = 8			9,667	interclasse	n = lignes
				F =	8,731		k = colonnes
				f =	2,947	p = 0,05	
					4,568	p = 0,01	
					8,735	p = 0,0003	

VAR / k	1,43	0,86	1,07	1,07	4,43		
VAR / k		k = 4			1,107	intraclasse	intergroupe
VAR / TOTAL (N)		N = 32			1,208		N = total
VAR pour n		n = 8			9,667	interclasse	n = lignes
			F =		8,731		k = colonne
			f =		2,947	p = 0,05	
					4,568	p = 0,01	
					8,735	p = 0,0003	

total	76	84	90	70	320		
Moyennes	9,5	10,5	11,25	8,75	10		
S(Xi - Xmoy)^2	0	0	0	0	0		
Moy - Moy/moy	-0,5	0,5	1,25	-1,25			
variance intraclasse			31 / 28 =	1,107	1,107	N - k = 28 ddl	rhô2 à 28 ddl
estimation 2	0,25	0,25	1,5625	1,5625	3,625		
moyenne					1,208		
variance interclasse			(3,625 x 8) / 3 =		9,667	k - 1 = 3 ddl	
d'où F			F =	9,667 / 1,107 =	8,731		

14. Liaison linéaire entre 3 variables

Soit 3 variables X, Y et Z

Calcul de Z' par la fonction TENDANCE. Avec la matrice : Z et XY.

Formule :

$$Z' = aX + bY + c$$

1. Calcul des coefficients a, b et c, en fonction des valeurs centrées :

$$a = \frac{\sum y^2 \cdot \sum xz - \sum xy \cdot \sum yz}{\sum x^2 \cdot \sum y^2 - (\sum xy)^2}$$

$$b = \frac{\sum x^2 \cdot \sum yz - \sum xy \cdot \sum xz}{\sum x^2 \cdot \sum y^2 - (\sum xy)^2}$$

$$c = \bar{Z} - a\bar{X} - b\bar{Y}$$

ou

$$c = \frac{\sum Z - a\sum X - b\sum Y}{n}$$

2. Calcul du coefficient de corrélation multiple :

$$R_{z.xy}^2 = \frac{a\sum xz + b\sum yz}{\sum z^2}$$

ou en fonction des 3 coefficients de corrélation déjà calculés :

$$R_{z.xy}^2 = \frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}$$

Cette formule permet de déterminer quel est des 3 coefficients celui qui s'adapte le mieux à l'ensemble des points observés.

Le coefficient de corrélation multiple est significatif à condition que le rapport :

$$F = \frac{n-3}{2} \frac{R^2}{1-R^2}$$

soit significativement supérieur à 1 avec $n_1 = 2$ et $n_2 = n - 3$ ddl.

3. Calcul des coefficients de corrélation partielle :

1) coefficient de corrélation partielle entre X et Z liés par Y :

$$r_{xz.y} = \frac{r_{xz} - r_{yx}r_{yz}}{\sqrt{1-r_{xy}^2}\sqrt{1-r_{xz}^2}}$$

2) coefficient de corrélation partielle entre Y et Z liés par X :

$$r_{yz.x} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{1-r_{xy}^2}\sqrt{1-r_{xz}^2}}$$

NB : les coefficients de corrélation ordinaires prennent l'appellation de coefficients de corrélation totale.

Ou alors on passe directement à l'analyse en composantes principales (ACP) et à l'analyse factorielle discriminante (AFD).

15. Test Z sur un échantillon : comparaison des moyennes

$$Z = \frac{\bar{X} - \mu}{\sigma_x^-}$$

erreur type ET :

$$ET = \sigma_x^- = \frac{\sigma_x}{\sqrt{n}}$$

n est le nombre de sujets de l'échantillon

z critique : loi.normale.standard.inverse (1- α = 0,95)

16. Test t d'un échantillon : comparaison des moyennes

$$t = \frac{\bar{X} - \mu}{\sigma_x^-}$$

erreur type ET :

$$ET = \sigma_x^- = \frac{\sigma_x}{\sqrt{n}}$$

n est le nombre d'observations de l'échantillon

z critique : loi.student.inverse (α ;ddl)

17. Test t de 2 échantillons indépendants : différence de 2 moyennes

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{S_{x_1-x_2}}$$

variance pondérée :

$$S^2 \text{ pondérée} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1 + (n_2 - 1))}$$

erreur type ET de la différence :

$$ET = S_{x-x} = \sqrt{\frac{S_{\text{pondérée}}^2}{n_1} + \frac{S_{\text{pondérée}}^2}{n_2}}$$

18. Test t d'échantillons appariés

$$t = \frac{D}{S_D}$$

erreur type ET de la différence de la différence moyenne :

$$ET = S_D = \frac{S_D}{\sqrt{n}}$$

n est le nombre total de paires d'observations

où $S_D = \text{Racine} (\Sigma(D - D_{\text{moy}})^2 / (n - 1)) :$

$$S_D = \sqrt{\frac{\Sigma(D - \bar{D})^2}{n - 1}}$$

t critique : loi.student.inverse(α ; 2(n-1))

Prendre la fonction correspondante

19. Test d'égalité des espérances : observations appariées

Test Z sur 2 échantillons indépendants : différence des moyennes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{x_1-x_2}}$$

erreur type ET de la différence des moyennes :

$$ET = \sigma_{x_1-x_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

z critique : {loi.normale.standar.inverse(1- α)}

Fonction : Test z de la différence significative minimale.

20. Exécution automatique du Test F de Fisher-Snédecor

Test exécuté automatiquement à la page 8 de la MACRO

Observation : Pour ce faire, recopier sur la Macro où l'on a effectué les calculs, les paramètres de calcul ajoutés à la page 8 de la Macro de base. Ce faisant, nous avons évité d'alourdir la programmation, puisque les formules peuvent être recopiées et les calculs automatiquement exécutés.

Estimation

Coller les formules suivantes :

- 1) en E7 : =E6^2/(E3-2)
- 2) en F1 : **F**
- 3) en F2 : =D6/E7
- 4) en F3 : =INVERSE.LOIF(0,05;1;(E3-2))
- 5) en F4: =LOIF(F2;1;(E3-2))
- 6) en G1: %
- 7) en G2: =E6
- 8) en G3: =1-E6
- 9) en H1: proportion
- 10) en H2: =E3*G2
- 11) en H3: =E3*G3

12) en H4: =SOMME(H2:H3)

13) en I2: Y non expliqué par X (éléments aberrants, atypiques ou indépendants)

14) en I3 : Y expliqué par X (hypothèse de liaison, si $r \geq 0,866 = \sqrt{3}/2 = \cos 30^\circ$)

15) en I4 : **Total**

On voit immédiatement ce qui se passe entre les 2 variables analysées.

Pour représenter la parabole des résidus, mettre:

1) en L8 : $\mathbf{Vr^2}$

2) en L9 : =I9^2

3) Puis cliquer sur le bouton en bas à droite de la cellule L9 pour étendre le calcul sur toute la longueur de la colonne.

4) Sélectionner les données de Vr et de Vr^2 pour faire la parabole.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

9

Exemple de Régression linéaire multiple

Le Petit Chaperon rouge

ACP et AFD

Densité, intensité, intention

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce neuvième chapitre, nous exécutons une approche d'analyse lexicale, textuelle et discursive du plus court des 8 contes de Perrault, *Le Petit Chaperon rouge*, que tout le monde connaît, et qui furent publiés sous le titre de *Contes de ma mère l'Oye*, en 1697. Ces contes connurent un succès immédiat. Peu après la mort de Charles Perrault le 16 mai 1703, la plupart de ces contes furent transcrits, par Grimm, ou mis en musique par Offenbach, Rossini, Bartok. Et Gustave Doré en fit une magnifique édition illustrée. Plus près de nous, Henri Galeron a continué selon la tradition...

Nous allons soumettre le texte au traitement analytique de STABLEX, non dans le but de le dessécher, mais au contraire d'en découvrir toutes les structures, d'en faire ressortir toutes les caractéristiques et d'en faire vibrer toutes les finesses que l'on ne perçoit pas toujours à la première lecture, et parfois même au terme de plusieurs lectures ou relectures.

Nous commencerons donc par présenter le texte du *Petit Chaperon rouge*, le résumer et le définir dans ses grandes lignes.

Mais pourquoi avoir choisi *Le Petit Chaperon rouge*, me direz-vous ? Pour des raisons toutes simples. D'abord, parce que c'est un texte court, agréable à lire, un conte connu et apprécié. Ensuite, parce qu'on croit le connaître, alors que l'analyse va nous montrer que l'on est bien souvent loin de compte. Enfin, parce qu'il va nous permettre, croyant le connaître, de mieux saisir la méthode d'analyse, de nous familiariser avec STABLEX et avec le traitement analytique que nous proposons.

Pour ce faire, nous présenterons dans un premier temps les graphiques qui mettent en évidence la place de ce deuxième conte dans le concert des huit contes en prose de *Ma mère l'Oye*, par le biais de l'ACP (l'Analyse en Composantes Principales). Dans un deuxième temps, nous irons directement à l'AFD (l'Analyse Factorielle Discriminante) pour en saisir toutes les subtilités, même si la longueur du chapitre nous oblige à procéder de façon synthétique. Enfin, nous irons tranquillement vers les relevés séquentiels, non dans le but de montrer comment faire des dictionnaires ou des lexiques, mais plus simplement, en restant dans le cadre de notre analyse, voir combien tous les éléments comptent dans une analyse, et qu'on ne peut surtout pas les couper de leur environnement textuel et encore moins bouleverser l'ordre naturel du texte, sous peine de tout dénaturer, et de prendre le texte pour « prétexte... à élucubration », ce qui serait tout à fait contraire à la méthode et à la méthodologie développée par STABLEX, et à l'esprit critique et scientifique.

1. Le texte, résumé et présentation

Voici le texte intégral du conte du Petit Chaperon rouge, que nous avons retranscrit à partir de l'œuvre de Gallimard, mais en changeant la présentation de façon à en faire ressortir au mieux les qualités narratives, notamment celles inhérentes au dialogue entre le Loup et le Chaperon rouge, là où se joue le drame qui fait la leçon et la morale du conte.

Le Petit Chaperon rouge.

Il était une fois une petite fille de Village, la plus jolie qu'on eût su voir ; sa mère en était folle, et sa mère-grand plus folle encore. Cette bonne femme lui fit faire un petit chaperon rouge, qui lui seyait si bien, que partout on l'appelait le Petit Chaperon rouge.

Un jour sa mère ayant cuit et fait des galettes, lui dit :

– Va voir comme se porte ta mère-grand, car on m'a dit qu'elle était malade, porte-lui une galette et ce petit pot de beurre.

Le Petit Chaperon rouge partit aussitôt pour aller chez sa mère-grand, qui demeurait dans un autre Village. En passant dans un bois elle rencontra compère le Loup, qui eut bien envie de la manger ; mais il n'osa, à cause de quelques bûcherons qui étaient dans la Forêt. Il lui demanda où elle allait ; la pauvre enfant qui ne savait pas qu'il est dangereux de s'arrêter à écouter un Loup, lui dit :

– Je vais voir ma Mère-grand, et lui porter une galette avec un petit pot de beurre que ma Mère lui envoie.

– Demeure-t-elle bien loin ? lui dit le Loup.

– Oh ! oui, dit le Petit Chaperon rouge, c'est par-delà le Moulin que vous voyez tout là-bas, là-bas, à la première maison du Village.

– Eh bien, dit le Loup, je veux l'aller voir aussi ; je m'y en vais par ce chemin ici, et toi par ce chemin-là, et nous verrons à qui plus tôt y sera.

Le Loup se mit à courir de toute sa force par le chemin qui était le plus court, et la petite fille s'en alla par le chemin le plus long, s'amusant à cueillir des noisettes, à courir après des papillons, et à faire des bouquets des petites fleurs qu'elle rencontrait.

Le Loup ne fut pas longtemps à arriver à la maison de la mère-grand ; il heurte : Toc, toc.

– Qui est là ?

– C’est votre fille, le Petit Chaperon rouge, dit le Loup, en contrefaisant sa voix, qui vous apporte une galette, et un petit pot de beurre que ma Mère vous envoie.

La bonne mère-grand qui était dans son lit à cause qu’elle se trouvait un peu mal, lui cria :

– Tire la chevillette, la bobinette cherra.

Le Loup tira la chevillette, et la porte s’ouvrit. Il se jeta sur la bonne femme, et la dévora en moins de rien ; car il y avait plus de trois jours qu’il n’avait mangé. Ensuite il ferma la porte, et s’alla coucher dans le lit de la mère-grand, en attendant le Petit Chaperon rouge, qui quelque temps après, vint heurter à la porte : Toc, toc.

– Qui est là ?

Le Petit Chaperon rouge, qui entendit la grosse voix du Loup, eut peur d’abord, mais croyant que sa mère-grand était enrhumée, répondit :

– C’est votre fille, le Petit Chaperon rouge, qui vous apporte une galette et un petit pot de beurre que ma Mère vous envoie.

Le Loup lui cria, en adoucissant un peu sa voix :

– Tire la chevillette, la bobinette cherra.

Le Petit Chaperon rouge tira la chevillette, et la porte s’ouvrit.

Le Loup, la voyant entrer, lui dit en se cachant dans le lit sous la couverture :

– Mets la galette et le petit pot de beurre sur la huche, et viens te coucher avec moi.

Le Petit Chaperon rouge se déshabille, et va se mettre dans le lit, où elle fut bien étonnée de voir comment sa mère-grand était faite en son déshabillé ; elle lui dit :

– Ma mère-grand, que vous avez de grands bras !

– C’est pour mieux t’embrasser, ma fille.

– Ma mère-grand, que vous avez de grandes jambes !

– C’est pour mieux courir, mon enfant.

– Ma mère-grand, que vous avez de grandes oreilles !

– C’est pour mieux écouter, mon enfant.

– Ma mère-grand, que vous avez de grands yeux !

– C’est pour mieux voir, mon enfant.

– Ma mère-grand, que vous avez de grandes dents !

– C’est pour te manger.

Et, en disant ces mots, ce méchant Loup se jeta sur le Petit Chaperon rouge, et la mangea.

MORALITÉ

On voit ici que de jeunes enfants,
Surtout de jeunes filles,
Belles, bien faites, et gentilles,
Font très mal d’écouter toute sorte de gens,
Et que ce n’est pas chose étrange,
S’il en est tant que le loup mange.
Je dis le loup, car tous les loups
Ne sont pas de la même sorte :
Il en est d’une humeur accorte,

**Sans bruit, sans fiel et sans courroux,
Qui privés, complaisants et doux,
Suivent les jeunes Demoiselles
Jusque dans les maisons, jusque dans les ruelles ;
Mais hélas ! qui ne sait que ces loups doucereux,
De tous les loups sont les plus dangereux.**

- *Résumé du conte.* Il est facile de résumer ce conte, que tout le monde connaît :

C'est l'histoire du méchant Loup, qui ruse autant qu'il peut, pour abuser de la jeune fille qu'il convoite, une jeune fille innocente et pure. Tant et si bien qu'à la fin, il mange le Petit Chaperon rouge, cette belle jeune fille de Village, pure et innocente, dont le seul tort a été de ne pas se méfier de ce Loup vicieux, « de ce Loup doucereux et dangereux », une métaphore qui est, hélas ! d'une éternelle actualité.

La leçon de morale est dite et faite. Il suffit de la lire à la fin du conte. C'est un avertissement aux parents et aux jeunes filles ingénues et sans malice, qui doivent se garder de fréquenter « ces loups doucereux, / (qui) De tous les loups sont les plus dangereux », comme il est dit dans la leçon de Morale, qui fait de ce conte un *exemplum*, au sens strict du terme.

C'est l'histoire d'un Loup qui veut manger une oie blanche. L'histoire d'un prédateur et d'une proie.

C'est la fable d'un Loup vicieux qui va dévorer le *Petit Chaperon rouge*, une jeune fille de Village, parce qu'elle est innocente et pure.

C'est le mythe de l'éternel retour, ou le retour de l'éternel mythe, un mythe éternel, qui se passe de tout commentaire.

- *Analyse logique du conte.*

L'analyse logique se fonde sur le rapport ordonné [/intention/ + /action/] pour définir à la fois [/l'acte/] et la responsabilité de son auteur. À savoir, la formule : [/intention/ + /action/ ⇔ /acte/] ayant pour corollaire la définition de la responsabilité de l'acte, en fonction des valeurs d'autodétermination qui la caractérisent. En fin de compte, on forme le jugement de valeur morale qui est un jugement de droit, et non un jugement de fait.

Ainsi, dans le *Petit Chaperon rouge*, on voit que les intentions des trois femmes (de la mère, de la grand-mère et de la petite fille, le « Petit Chaperon rouge ») sont des intentions saines : elles sont généreuses et sans arrière-pensées. Aussi leur comportement est-il purement et simplement d'une haute qualité morale.

En revanche, pour ce qui est du Loup (pris métaphoriquement pour l'homme « pervers » ou « vicieux »), il en va tout autrement. Les intentions qu'il affiche sont des intentions pernicieuses, camouflées, il évite surtout de se démasquer trop vite, et de se faire surprendre par les bûcherons ou de se faire remarquer par « la petite fille de Village ». Aussi ruse-t-il jusqu'à tromper d'abord la « mère-grand », puis « le Petit Chaperon rouge » qu'il dévore l'une après l'autre. Du point de vue logique, les intentions affichées sont à l'inverse de ce qu'elles sont dans la réalité, laquelle est dévoilée par l'action de « manger » les deux femmes.

Bref, d'un point de vue logique et aussi métaphorique, le thème du conte est dans l'esprit de l'adage latin qui dit : *homo homini lupus est*, l'homme est un loup pour l'homme.

L'analyse logique ne peut être remise en cause par l'analyse lexicale, textuelle et discursive faite au moyen du STABLEX. Au contraire, elle doit être confortée, précisée et

détaillée, dans la mesure où elle va dévoiler avec minutie les qualités techniques d'écriture et de composition qui font le charme du conte.

2. L'ACP (Analyse en Composantes Principales)

L'ACP est une technique graphique qui consiste à configurer la position des variables dans un nuage de points en fonction des distances qui les séparent, lesquelles distances préfigurent les liaisons structurales des données.

C'est ainsi que l'ACP ouvre la voie à l'AFD (Analyse Factorielle Discriminante), qui représente l'analyse qualitative par excellence.

C'est à ce titre que les huit contes en prose de Perrault sont appelés T1, T2, T3... T8, pour désigner les Huit « *Contes de ma mère l'Oye* » :

T1 La Belle au bois dormant

T2 Le Petit Chaperon rouge

T3 La Barbe-bleue

T4 La Maître Chat, ou le Chat botté

T5 Les fées

T6 Cendrillon, ou la Petite Pantoufle de verre

T7 Riquet à la Houppe

T8 Le Petit Poucet

Voyons d'abord les données statistiques de base résumées dans le tableau suivant :

	Total	T1	T2	T3	T4	T5	T6	T7	T8
Occurrences	18094	3618	769	2002	1742	940	2509	2704	3810
Vocables	2687	969	271	622	543	347	714	726	951
Hapax	1484	341	42	156	142	61	198	221	323
Répétition	0,4477	0,6481	0,8450	0,7492	0,7385	0,8242	0,7227	0,6956	0,6604

Dans la première ligne figure le nombre total d'occurrences de chaque variable. Dans la deuxième, figure le nombre total de vocables. Dans la troisième, le nombre d'hapax (ou vocables de fréquence 1 propres à chaque variable) ; et, dans la quatrième, en fonction du nombre d'hapax, le taux de répétition propre à chaque variable.

Au vu des résultats de l'analyse fournis par STABLEX, on voit que le texte T2 du *Petit Chaperon rouge* tient une place à part dans le corpus : c'est le texte le plus court, mais qui a le plus fort taux de répétition.

Les données statistiques le mettent immédiatement en exergue.

Est-ce pour autant un texte singulier, différent des autres ? Un texte de facture différente ou d'une autre plume ?

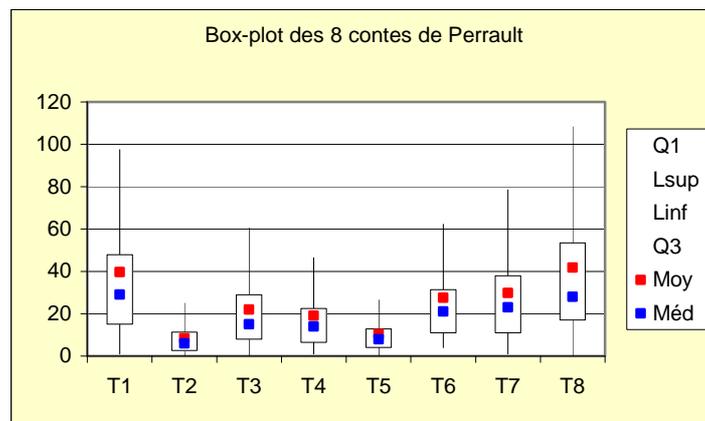
Compte tenu de ce que nous savons du soin apporté par Perrault à la publication de ses contes, personne n'oserait contester la paternité et l'authenticité des 8 contes de « *ma mère l'Oye* » : ils sont bel et bien de la même veine et de la même plume.

Néanmoins, pour nous en assurer, nous soumettons le corpus aux différents tests que nous propose la statistique, et qui sont effectués par la MACRO associée à STABLEX. Nous pouvons ainsi contrôler toute la chaîne d'analyse et d'étude. À la base, il y a les textes et les lexiques, qui donnent lieu aux 2 tables de contingence fondamentales dans la description statistique : la TDF (Table de Distribution des Fréquences) et la TDR (Table des Écartés Centrés Réduits ou des densités). Il y a ensuite l'élaboration des données qui ouvrent la voie,

entre autres, à l'ACP (Analyse en Composantes Principales, c'est-à-dire des variables considérées dans leur intégralité) et l'AFD (Analyse Factorielle Discriminante, c'est-à-dire des éléments constitutifs de la variable). Le tout accompagné des graphiques descriptifs qui permettent de visualiser les phénomènes inhérents au corpus et aux variables qui le constituent : le tout s'explique par la partie et la partie par le tout.

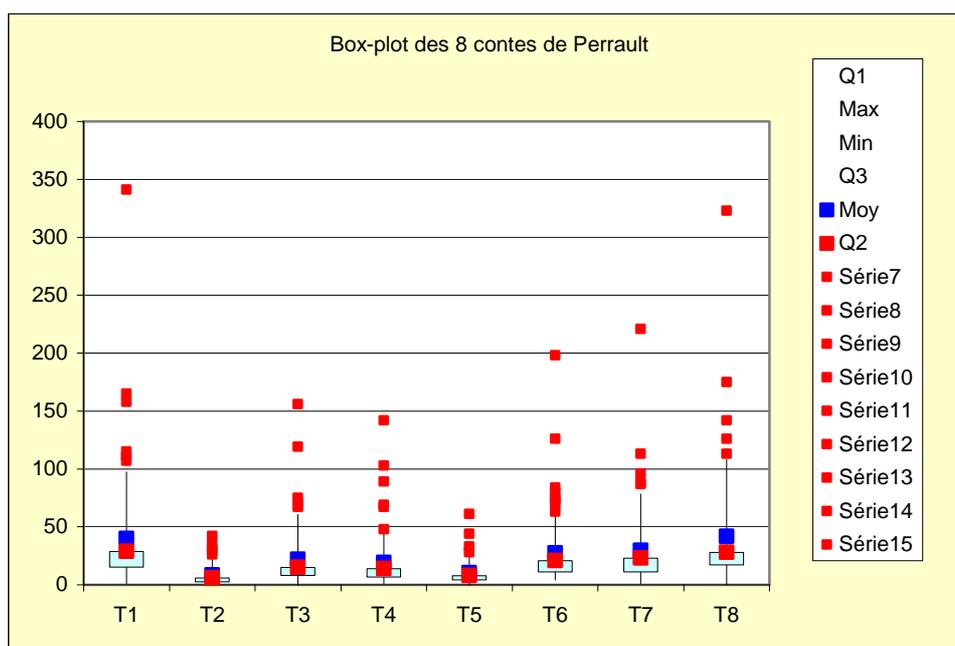
Le *corpus* est normalement distribué au vu du test du χ^2 de Fisher qui affiche une valeur de 100%. Ce qui signifie que les contes sont de la même veine. Chose dont personne n'a jamais douté, et qu'il serait en effet difficile de contester ou de remettre en cause.

Toutefois, pour répondre à la première question qui vient à l'esprit, de savoir si le texte T2 du *Petit Chaperon rouge* appartient bien au groupe des 8 contes, nous observons le graphique des boîtes à moustaches (ou la *Box-plot* de Tukey, en termes techniques) :



Les « boîtes à moustaches » sont identiques pour toutes les variables, mais elles affichent des valeurs propres qui préfigurent des dispersions propres à chacune d'elles : asymétrie, dispersion, étendue des distributions. On voit notamment que la moyenne et la médiane se projettent toujours de la même façon dans l'interquartile : la moyenne tend toujours vers le 3^{ème} quartile (le plus grand). Mais on voit aussi que la variable T2 du *Petit Chaperon rouge* est resserrée sur elle-même, suivant une forte concentration des éléments lexicaux.

Mais on peut encore mettre en évidence les valeurs « aberrantes » de chaque variable :

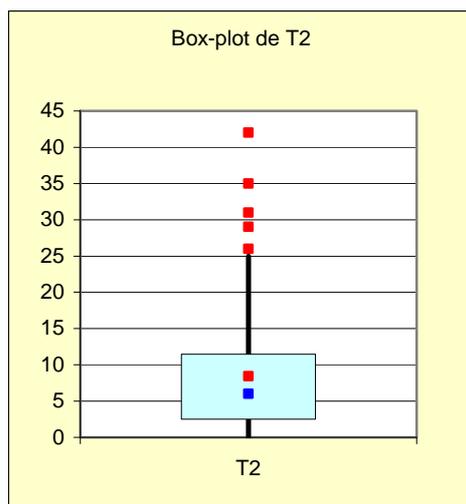


Ce graphique met en évidence les éléments remarquables, atypiques ou déviants, de chaque variable à cause de leur valeur hautement significative¹ : on peut ainsi dénombrer les éléments des « *outliers* » et voir que T1, T3 et T4 en comptent 6 chacun ; que T2, T5, T7 et T8 n'en comptent que 5 ; mais que T6 en revanche en compte 9. La comparaison est immédiate : on en a une vision synthétique, c'est le but recherché.

Voilà un ensemble de traits génériques qui n'apparaissent certainement pas à première vue ni à la première lecture, notamment en ce qui concerne le conte T2 du *Petit Chaperon rouge*.

¹ Les graphiques ne sont là que pour montrer les qualités des dispersions. Néanmoins, pour affiner l'analyse et cibler avec précision les valeurs des *outliers*, il vaut mieux se rapporter aux valeurs centrées réduites fournies par la TDR, sachant que les limites supérieure et inférieure des barrières correspondent sensiblement au $z = \pm 2,7$ des valeurs centrées réduites. Alors, comme il est aisé de faire varier les limites supérieure et inférieure des valeurs centrées réduites dans la MACRO, il est facile de repérer, et avec quelle précision, les « valeurs aberrantes ou atypiques » de la matrice, et donc d'identifier les éléments à forte variation statistique (hautement significatifs). Tout cela est expliqué dans les ouvrages de référence.

NB : Les valeurs qui servent à former les graphiques sont certes des valeurs réelles, puisque le corpus est parfaitement défini, et qu'il s'agit d'une statistique descriptive, et non d'une « statistique hypothétique » qui travaille sur des échantillons, mais, ce faisant, les statisticiens travaillent de façon tautologique en créant des « variables artificielles » pour des raisons tout aussi artificielles... d'où on ne retire qui n'y ait été artificiellement introduit.



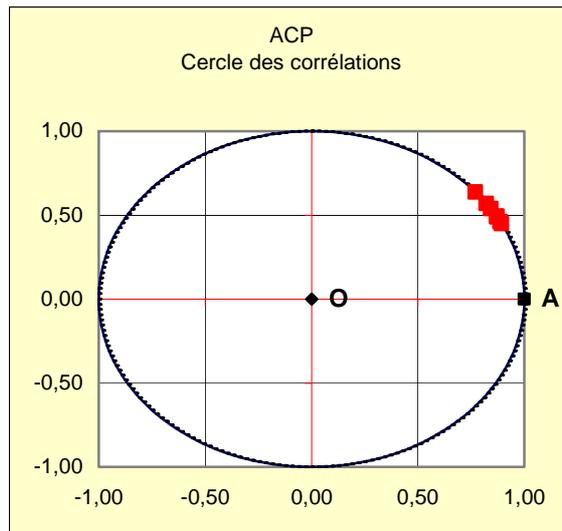
Aussi convient-il de pousser l'analyse et l'observation pour en découvrir les raisons.

Certes, on pourrait aller d'entrée de jeu vers les valeurs centrées réduites² et considérer les éléments lexicaux qui font la singularité de cette variable par rapport aux autres, mais, rappelons-le, nous laissons présentement de côté les descriptions statistiques de base concernant l'analyse lexicale, textuelle et discursive (dont on trouvera la description dans les différents ouvrages de référence), pour nous tourner vers l'ACP et l'AFD, qui vont retenir toute notre attention.

Voyons d'abord les qualités et les caractéristiques de T2 par le biais de l'ACP, sans toutefois entrer dans tous les détails techniques qui sont décrits dans les ouvrages techniques qui accompagnent le soft STABLEX.

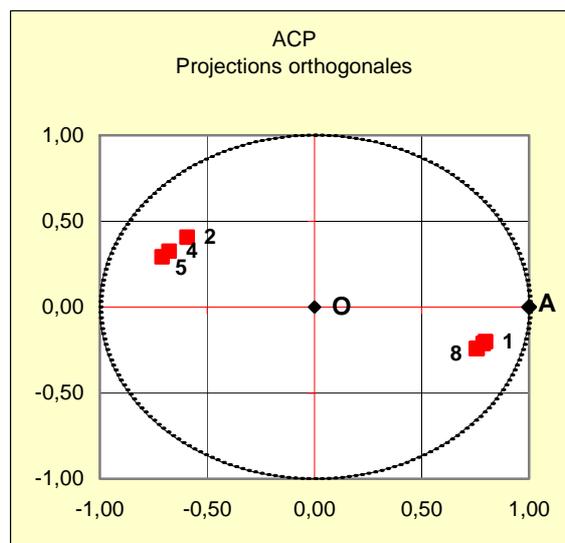
Le graphique en nuage de points confirme que les 8 variables sont étroitement liées, bien que l'une d'elles semble se détacher du groupe, comme il apparaît dans le premier graphique ci-après :

² Il convient de souligner sans ambiguïté et sans ambages que le calcul de la *box-plot* se fait sur des lieux géométriques (médiane et quartiles), des « nombres sans dimension », alors que les calculs algébriques (de la TDR, des densités, des corrélations...) sont des calculs matriciels de mesure, de comparaison et d'intégration. On évitera donc de confondre la médiane, et les quartiles en tant que lieux géométriques, avec la moyenne et l'écart-type, en tant que références algébriques. La médiane n'est qu'un indicateur de position (un lieu géométrique), insensible aux valeurs extrêmes, alors que la moyenne est une valeur centrale dans le calcul algébrique, calcul par excellence d'intégration et de comparaison.



On voit que les 8 contes sont regroupés dans un même espace restreint du cercle de centre O et de rayon 1.

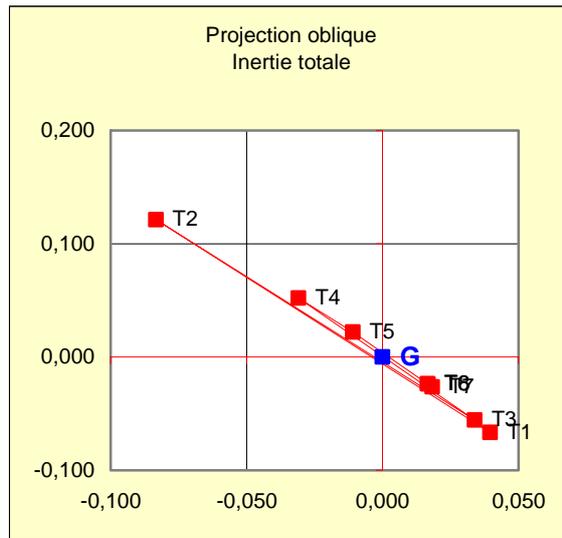
Voir également le cercle des projections orthogonales ($\cos^2 + \sin^2 = 1$)³ :



Poussons encore plus loin l'observation du corpus, en examinant à la loupe la distribution des 8 variables autour du centre de gravité (dans une inertie totale) afin d'avoir une vision et une idée exacte des distances qui les séparent, telles qu'on peut les observer à travers une projection oblique, une projection orthogonale, un dendrogramme et une distribution isotrope, par exemple.

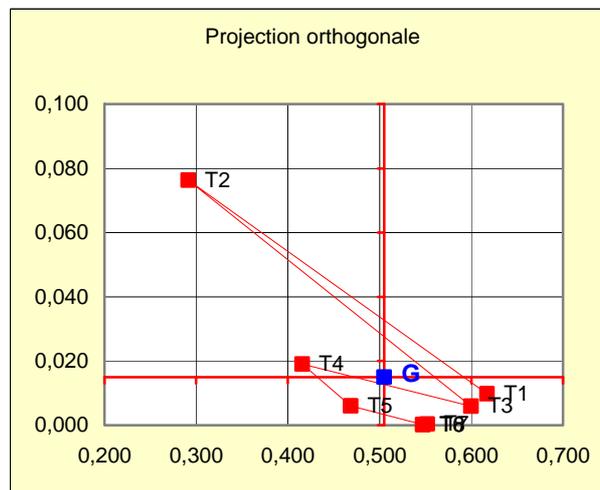
- La projection oblique, qui élargit le champ de vision, montre clairement la *position* de T2, le *Petit Chaperon rouge*, par rapport aux 7 autres contes du corpus, tels qu'ils se distribuent autour du centre de gravité dans une inertie totale :

³ La projection orthogonale se fait à partir du coefficient de détermination (le carré du cosinus et du sinus) qui indique la part de variance de Y expliquée en fonction de celle, supposée connue, de X. Les carrés sont affectés du signe respectif de r'' et p'' de l'inertie totale (valeur calculée/valeur absolue).

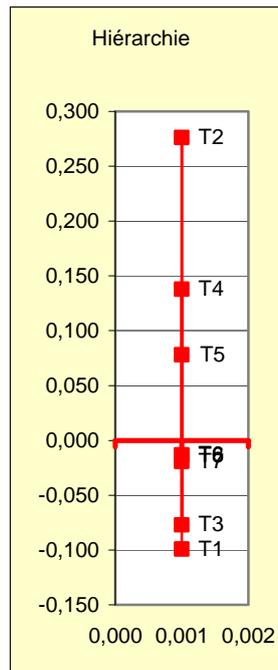


Les valeurs de calcul ne sont pas des valeurs de mesure et de dimension, mais des valeurs de positionnement, ce qui doit nous permettre de bien saisir la particularité de l'ACP (où les points se distribuent autour du centre de gravité G (0 ; 0)).

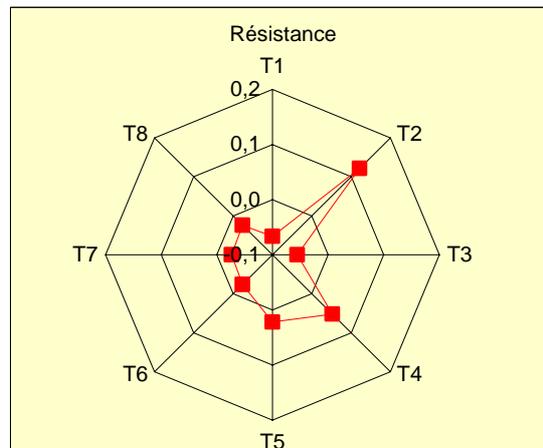
- La projection orthogonale accentue encore d'avantage l'éloignement de T2 qui est rejeté dans la partie haute à gauche du graphique, mettant le *Petit Chaperon rouge* dans une position à part par rapport aux 7 autres contes qui se situent tous (sensiblement) en dessous de l'axe des abscisses passant par le centre de gravité G (0,505 ; 0,015).



- Le dendrogramme ci-après confirme cette position haute de T2 dans la « hiérarchie structurale » des 8 contes (3 contes au-dessus de la moyenne, T2, T4 et T5, et 5 en dessous, avec T3 et T1 en bas de l'échelle) :



• La distribution des vecteurs isotropes⁴ des 8 variables montre combien T2 est projeté par une force centrifuge à l'extérieur du groupe, lui conférant ainsi une place à part dans le concert des 8 contes :



Sachant que le texte du *Petit Chaperon rouge* est le texte le plus court, toutes les questions que l'on est en droit de se poser à partir des graphiques précédents, vont trouver une raison, une explication et une justification dans l'AFD (l'Analyse Factorielle Discriminante) qui suit.

3. L'AFD (Analyse Factorielle Discriminante)

L'Analyse Factorielle Discriminante (AFD) n'est rien moins qu'une régression multiple réduite à deux éléments : la variable estimée Y comme « *critère* » et la variable explicative X comme « *prédicteur* ». Bref, pour faire court, disons que l'estimation, bien qu'ayant l'allure

⁴ Voir Cap. 8 du *Livre de statistique*, publié en 2006. Disponible sur CD.

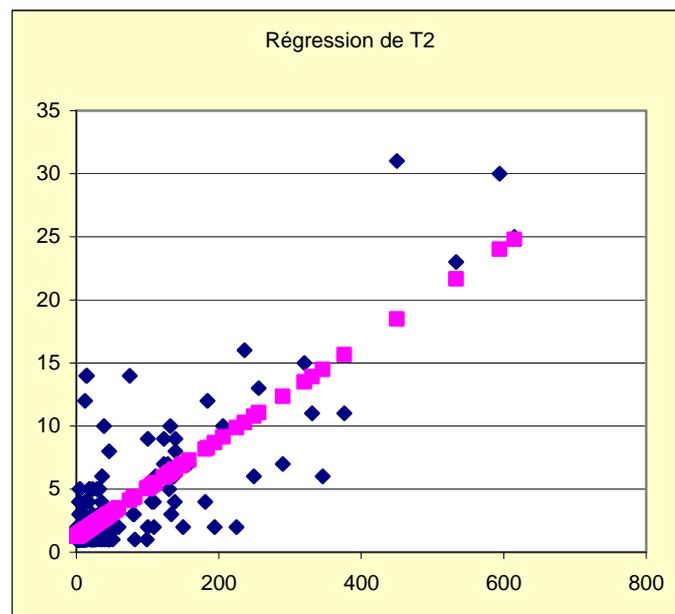
d'une régression multiple, n'est en réalité qu'une régression simple, que la MACRO traite automatiquement.

L'estimation met en évidence un certain nombre de caractères d'une importance fondamentale pour une analyse factorielle discriminante :

1. le nombre de facteurs, qui est de 271 vocables pour T2
2. l'indice de corrélation est moyen $r = 0,822$ ($< 0,866 = \sqrt{3}/2 = \cos 30^\circ$)
3. l'indice de détermination, qui vaut $r^2 = 0,676$
4. la variabilité de Y par rapport à X vaut 0,431
5. le pourcentage de facteurs de Y liés à X est donc de 43,1%
6. la proportion d'éléments de Y liés à X est de 117 / 271
7. le pourcentage de facteurs de Y indépendants de X est de 56,9%
8. la proportion d'éléments de Y non liés à X est de 154 / 271
9. le taux de facteurs ayant un khi2 hautement significatif : 10,33%
10. l'indice du seuil de signification de dt vaut 0,384
11. le test F est hautement significatif : $F = 560,681 < f_{(0,05; 1; 269)} = 3,876$.
12. d'où l'acceptation de l'hypothèse H_1 : T2 se détache nettement du corpus.

Retenons, en gros, que l'hypothèse d'une relation linéaire entre T2 et le corpus des 8 contes permet d'expliquer à concurrence de 67,6% la variance du *Petit Chaperon rouge*, avec une variabilité de T2 de l'ordre 43,1% par rapport à l'ensemble : d'où les 117 vocables sur les 271.

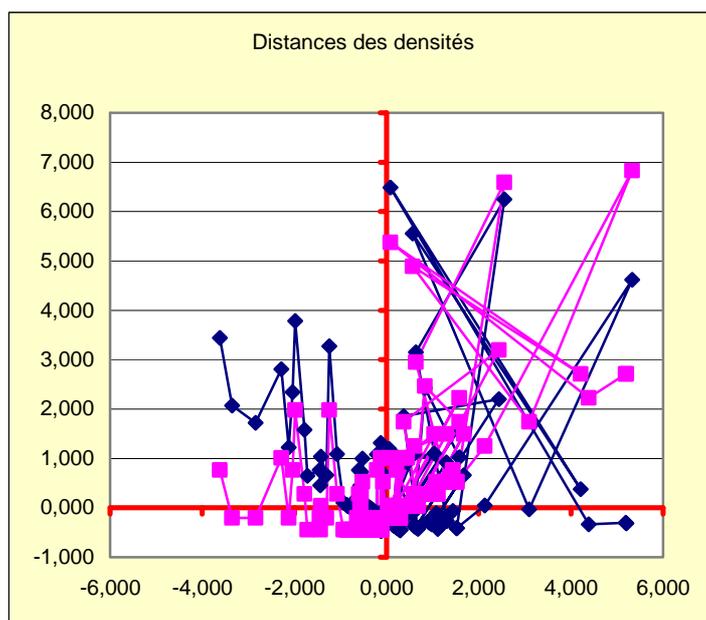
C'est au vu de toutes ces indications d'une grande cohérence que l'analyse peut être engagée à coup sûr : elle doit permettre d'identifier, d'observer, d'expliquer, de justifier et de vérifier les raisons qui font que tous ces éléments sont déterminés et/ou prédéterminés, que certains occupent une place de choix, voire prépondérante, dans la composition, dans l'écriture, dans la logique, dans le raisonnement, dans l'esprit du conte, pour ne pas dire dans l'esprit de l'auteur.



La distribution des éléments (en bleu) autour de la droite de régression (en rouge) montre bien que T2 offre une certaine résistance à l'intégration parfaite.

Pour avoir une idée exacte du contenu lexical du conte, nous donnons en annexe la page complète de l'estimation de T2, *Le Petit Chaperon rouge*, en fonction des « 8 Contes de ma mère l'Oye ».

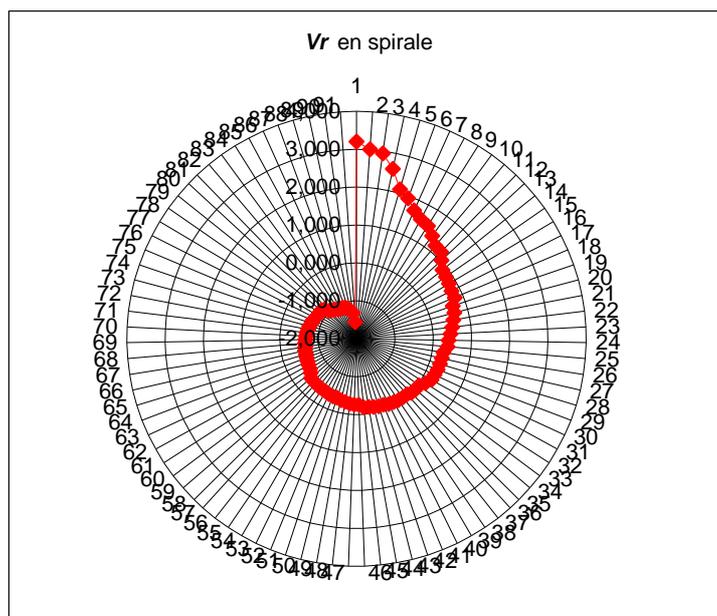
Néanmoins, à titre d'exemple et pour plus ample aperçu, voyons le nuage des points des « moindres carrés » des densités qui illustrent nettement la qualité de la distribution de T2 par rapport à l'ensemble des données du corpus : en fonction des valeurs centrées réduites, on repère immédiatement les éléments dits « aberrants », « remarquables », « atypiques », ou encore « à forte valeur thématique ou grammaticale », ils sont parfaitement identifiés.



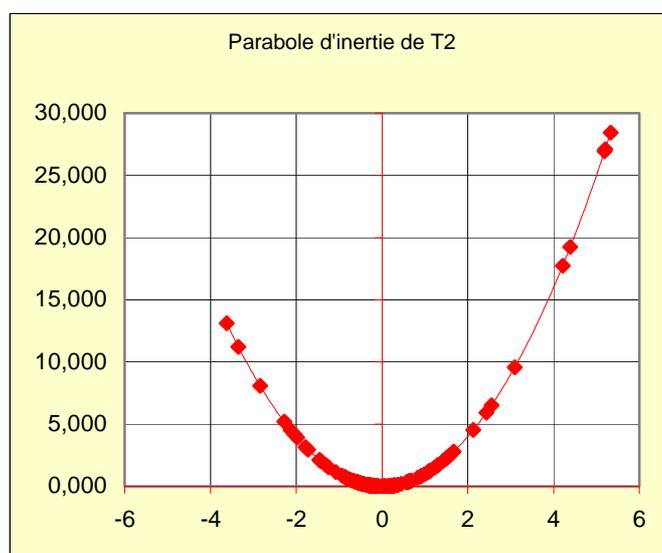
Plus les distances sont grandes et plus les résistances à l'intégration sont fortes. Ces résistances sont positives ou négatives en fonction de la distance verticale qui sépare l'ordonnée de chaque facteur (en bleu, favorable au *critère Y* (T2), et en rouge, favorable au *prédicteur X*).

Les valeurs centrées réduites (de z ou de Vr) mettent en évidence les choix (préférentiels, thématiques, logiques, grammaticaux...) qui appartiennent aux motivations et au style de Perrault au moment où il a écrit et composé (*cum-ponere = mettre ensemble, ajuster, assembler*) *Le Petit Chaperon rouge*.

Voici le spectre isotrope sous forme de spirale qui s'en dégage :



Néanmoins, pour faciliter la lecture lexicale et l'initiation à la méthode STABLEX, nous allons examiner les éléments constitutifs du lexique en fonction des deux branches de la parabole d'inertie qui en donne la configuration complète :



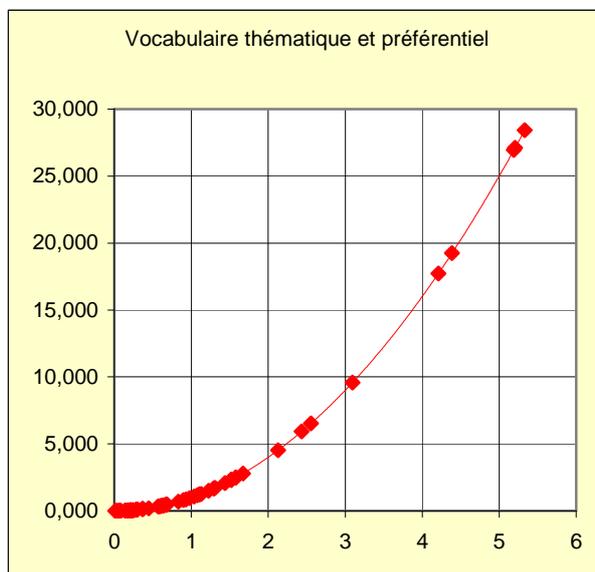
Sur la branche positive figurent les éléments caractéristiques qui font l'objet d'un choix préférentiel ou thématique, d'une logique de composition, ou de tout ce qui fait l'art et la manière d'écriture du *Petit Chaperon rouge*.

Sur la branche négative figurent, au contraire, les éléments de base, « communs » à l'écriture des contes et propres au style de Perrault.

Observons le vocabulaire porté par les deux branches de la parabole en nous laissant guider par les valeurs des densités (des résidus Vr et des écarts centrés réduits z) et des distances dt du khi2).

3.1 Le vocabulaire thématique et le vocabulaire de prédilection portés par la branche positive de la parabole

Paramètres de sélection : d'abord on sélectionne les valeurs positives de $V_r \geq 0$ que l'on range suivant l'ordre décroissant de dt .



La branche positive de la parabole porte les 79 vocables que voici, classés par ordre décroissant de dt .

Mot	Occ	T2	z	Vt	Vr	Vc	dt	Vr ²
Loup	14	14	17,760	2,709	5,207	-0,312	3,152	27,113
Mère-grand	15	14	17,103	2,709	5,191	-0,300	3,138	26,943
Petit Chaperon rouge	12	12	16,442	2,223	4,388	-0,334	2,685	19,251
de	615	25	-0,227	5,377	0,082	6,484	2,563	0,007
est	75	14	6,189	2,709	4,211	0,378	2,296	17,734
et	533	23	0,075	4,892	0,569	5,557	1,907	0,323
ma	39	10	6,622	1,738	3,095	-0,029	1,811	9,578
le	450	31	2,775	6,832	5,332	4,618	1,240	28,427
c'	46	8	4,418	1,253	2,129	0,050	1,217	4,531
beurre	5	5	10,613	0,525	1,520	-0,414	1,050	2,310
galette	5	5	10,613	0,525	1,520	-0,414	1,050	2,310
pot	5	5	10,613	0,525	1,520	-0,414	1,050	2,310
la	594	30	0,967	6,590	2,555	6,246	1,030	6,528
que	320	15	0,388	2,951	0,638	3,148	0,897	0,407
porte	18	5	4,948	0,525	1,308	-0,267	0,867	1,710
voir	36	6	3,693	0,768	1,440	-0,063	0,862	2,073
avez	19	5	4,768	0,525	1,291	-0,255	0,853	1,667
chevillette	4	4	9,493	0,283	1,110	-0,425	0,816	1,232
Toc	4	4	9,493	0,283	1,110	-0,425	0,816	1,232
petit	23	5	4,158	0,525	1,226	-0,210	0,797	1,503
fille	29	5	3,468	0,525	1,128	-0,142	0,713	1,272
dit	100	9	2,355	1,496	1,673	0,661	0,706	2,799
enfant	12	4	4,994	0,283	0,980	-0,334	0,704	0,959

<i>mère</i>	32	5	3,190	0,525	1,079	-0,108	0,671	1,164
<i>mieux</i>	15	4	4,304	0,283	0,931	-0,300	0,662	0,866
<i>lit</i>	17	4	3,941	0,283	0,898	-0,278	0,634	0,806
<i>envoie</i>	4	3	7,014	0,040	0,684	-0,425	0,569	0,468
<i>courir</i>	5	3	6,180	0,040	0,668	-0,414	0,555	0,446
<i>écouter</i>	5	3	6,180	0,040	0,668	-0,414	0,555	0,446
<i>village</i>	5	3	6,180	0,040	0,668	-0,414	0,555	0,446
<i>voix</i>	5	3	6,180	0,040	0,668	-0,414	0,555	0,446
<i>un</i>	206	10	0,430	1,738	0,369	1,859	0,534	0,136
<i>qui</i>	236	16	1,926	3,194	2,435	2,198	0,530	5,929
<i>grandes</i>	7	3	5,064	0,040	0,635	-0,391	0,527	0,403
<i>jeunes</i>	7	3	5,064	0,040	0,635	-0,391	0,527	0,403
<i>loups</i>	7	3	5,064	0,040	0,635	-0,391	0,527	0,403
<i>vous</i>	132	10	1,894	1,738	1,577	1,022	0,505	2,486
<i>en</i>	256	13	0,657	2,466	0,830	2,425	0,493	0,690
<i>chemin</i>	10	3	4,037	0,040	0,586	-0,357	0,484	0,344
<i>était</i>	147	7	0,308	1,011	0,054	1,192	0,449	0,003
<i>dans</i>	123	9	1,686	1,496	1,298	0,921	0,384	1,684
<i>par</i>	35	4	2,105	0,283	0,604	-0,074	0,381	0,365
<i>apporte</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>bobinette</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>cherra</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>dangereux</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>jusque</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>là-bas</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>tire</i>	2	2	6,713	-0,202	0,291	-0,447	0,349	0,085
<i>cria</i>	3	2	5,359	-0,202	0,274	-0,436	0,335	0,075
<i>folle</i>	5	2	3,963	-0,202	0,242	-0,414	0,307	0,058
<i>ouvrit</i>	5	2	3,963	-0,202	0,242	-0,414	0,307	0,058
<i>mon</i>	23	3	2,091	0,040	0,374	-0,210	0,302	0,140
<i>vais</i>	6	2	3,531	-0,202	0,225	-0,402	0,293	0,051
<i>ici</i>	7	2	3,190	-0,202	0,209	-0,391	0,279	0,044
<i>jeta</i>	7	2	3,190	-0,202	0,209	-0,391	0,279	0,044
<i>te</i>	7	2	3,190	-0,202	0,209	-0,391	0,279	0,044
<i>tira</i>	7	2	3,190	-0,202	0,209	-0,391	0,279	0,044
<i>lui</i>	184	12	1,528	2,223	1,580	1,610	0,270	2,496
<i>cause</i>	9	2	2,673	-0,202	0,176	-0,368	0,251	0,031
<i>coucher</i>	9	2	2,673	-0,202	0,176	-0,368	0,251	0,031
<i>mal</i>	9	2	2,673	-0,202	0,176	-0,368	0,251	0,031
<i>sorte</i>	9	2	2,673	-0,202	0,176	-0,368	0,251	0,031
<i>va</i>	9	2	2,673	-0,202	0,176	-0,368	0,251	0,031
<i>grands</i>	11	2	2,291	-0,202	0,144	-0,346	0,223	0,021
<i>sont</i>	11	2	2,291	-0,202	0,144	-0,346	0,223	0,021
<i>bien</i>	111	6	0,603	0,768	0,215	0,785	0,191	0,046
<i>bonne</i>	31	3	1,498	0,040	0,243	-0,120	0,190	0,059
<i>plus</i>	128	7	0,684	1,011	0,364	0,977	0,182	0,132
<i>sa</i>	139	9	1,300	1,496	1,036	1,102	0,159	1,074
<i>manger</i>	16	2	1,636	-0,202	0,062	-0,289	0,153	0,004
<i>votre</i>	16	2	1,636	-0,202	0,062	-0,289	0,153	0,004
<i>maison</i>	18	2	1,443	-0,202	0,030	-0,267	0,124	0,001
<i>s'</i>	123	7	0,792	1,011	0,446	0,921	0,112	0,198
<i>petite</i>	19	2	1,356	-0,202	0,013	-0,255	0,110	0,000

<i>se</i>	139	8	0,880	1,253	0,610	1,102	0,089	0,373
<i>car</i>	41	3	0,974	0,040	0,080	-0,006	0,049	0,006
<i>sans</i>	41	3	0,974	0,040	0,080	-0,006	0,049	0,006
<i>sur</i>	44	3	0,844	0,040	0,031	0,027	0,007	0,001

Ce rangement appelle quelques remarques préliminaires⁵ :

1. les valeurs de *dt* rangées par ordre décroissant réorganisent les valeurs des résidus de *Vr* (les densités de prédilection) et les valeurs des écarts réduits centrés de *z* (les densités absolues), pour donner le nouvel ordre des vocables caractéristiques de T2
2. les vocables sont ainsi regroupés par strates : en tête, les éléments à forte dominante thématique (13 vocables en gras ayant un $dt \geq 1$, en rouge) ; puis, les vocables de la sous-dominante thématique (24 éléments en gras et en italique, ayant un $0,500 \leq dt < 1$), et enfin les éléments de développement et d'enrichissement du discours (ayant un $0 \leq dt < 0,500$). Soit au total 79 vocables, dont la trajectoire se projette dans la partie supérieure du cône isotrope.

Faisons un rapide commentaire pour souligner combien le chemin tracé par les valeurs statistiques est hautement significatif de la structure du conte, tous compartiments et tous registres d'écriture confondus.

a) Les 13 éléments fondamentaux de la « dominante thématique »

En tête figurent les trois personnages qui forment l'intrigue du conte : le **Loup**, la **Mère-grand** et le **Petit Chaperon rouge**. Les valeurs des densités de *z* et de *Vr* ne laissent aucun doute à ce sujet.

Le **Loup** ($z = 17,760$) est bien le protagoniste, suivi de la **Mère-grand** ($z = 17,103$), bien qu'il faille noter ici l'ambiguïté du personnage partiellement doublé par le **Loup** qui se glisse dans le lit de la grand-mère surtout pour dévorer la petite fille. Enfin vient le **Petit Chaperon rouge** ($z = 16,442$).

Le « relevé chronologique » des séquences donne la juste mesure du rôle de chaque personnage :

- Le **Loup** est un élément actif, un agent, à la fois sujet grammatical et sujet thématique. C'est un Loup qui parle, qui écoute, qui agit, qui manœuvre, qui manipule, qui ruse, qui trompe ; qui « dévore » et qui « mange ». Bref, c'est le Loup qui mène la danse⁶. Aussi est-ce sur lui qu'est braqué le projecteur de l'observation.

- La **Mère-grand** n'est qu'un faire-valoir pour le Loup. C'est le stratagème dont se sert Perrault pour ourdir la trame du conte. Le rôle de la grand-mère est de prime abord limité, puisqu'elle est dévorée en moins de deux, mais en fait décuplé, puisqu'il permet au Loup de

⁵ L'algèbre, selon Descartes, est la clé des autres sciences. C'est un instrument de mesure, d'estimation et de comparaison.

⁶ *Mener la danse* = être responsable de l'action ou l'instigateur. Le Loup va mener son action à sa guise, en rusant et trompant son monde. L'hypertexte créé par les densités en fait le protagoniste : toutes les densités convergent, celle de *z*, celle de *Vr* et celle de *dt* ; elles le placent toutes en tête de liste. D'où l'expression employée pour le caractériser : « car il y avait plus de trois qu'il n'avait pas mangé », nous dit le texte. Parallèlement, les expressions suivantes viennent immédiatement à l'esprit : « *Avancer à pas de loup* », c'est-à-dire « de manière silencieuse et sournoise » ; « *Approcher à pas de loup* », c'est-à-dire « avec détermination, de façon résolue » ; « *Avoir une faim de loup* », c'est-à-dire « être affamé et vorace ».

se glisser dans sa peau et dans son lit, en lui fournissant les clés du piège qui va se refermer sur la petite fille.

- Le *Petit Chaperon rouge* est alors la victime expiatoire, ou propitiatoire.

Après les trois personnages viennent trois substantifs – *beurre, galette, pot* – qui sont l'élément moteur de la tragédie : ils constituent le « mobile », non pas du crime, mais de l'exposition dangereuse qu'encourt la petite fille. Trois éléments clés que le Loup va exploiter à son avantage, en y associant les « clés » qui lui permettent de s'introduire dans la maison de la Mère-grand et de se glisser dans son lit pour « manger » le *Petit Chaperon rouge*. Ce sont les quatre vocables du chiasme qui figurent dans la sous-dominante :

– *Tire la chevillette, la bobinette cherra.*

Le Loup a bien retenu la leçon. Le Petit Chaperon rouge lui a fourni la première clé qui lui permet de s'introduire chez la Mère-grand :

- *Toc toc. « Qui est là ?*
- *C'est votre fille, le Petit Chaperon rouge (dit le Loup en contrefaisant sa voix), qui vous apporte une galette, et un petit pot de beurre que ma Mère vous envoie ».*

Ce faisant, le Loup obtient par la ruse la deuxième clé qui va lui permettre de faire tomber le Petit Chaperon rouge dans ses bras :

– *Tire la chevillette, la bobinette cherra.*

On peut dire que le Loup a pu manœuvrer à sa guise, et « mettre la pauvre enfant dans de beaux draps »⁷ ou « dans de sales draps » en la « fourrant dans son lit »⁸ pour la « manger »⁹.

L'importance du vocabulaire de fréquence 2 est de livrer au lecteur la clé du succès du Loup qui, par un tour de passe-passe, va dévorer successivement la Mère-grand et le Petit Chaperon rouge, dès l'instant qu'il s'est glissé dans le lit de la Mère-grand.

Ce vocabulaire de fréquence 2 augmente le rythme et la densité du discours, tenant ainsi le lecteur en haleine.

Parmi les mots grammaticaux qui figurent dans la zone de la thématique dominante, on remarque, au-delà des vocables de fréquence 2, la préposition *de* et la conjonction *et*, ou encore, dans la panoplie grammaticale qui entre dans le jeu du style direct, des vocables d'apparence tout aussi banale, *c'est, le, la, ma, que, avez, mieux...* mais qui jouent un rôle fondamental dans l'élaboration du discours :

⁷ Signifie « la mettre dans une situation critique ».

⁸ « *Mettre dans son lit ou jeter dans son lit* » signifie « avoir des rapports sexuels illicites » par allusion à la pauvre enfant qui est « déflorée ». C'est toute la symbolique du conte qu'il faudrait encore commenter.

⁹ Nous utilisons à dessein des expressions, devenues légendaires ou proverbiales, parce qu'elles sont dans l'esprit conte qui est tourné vers la leçon de morale, une leçon de vie ou de simple bon sens. Parmi ces expressions retenons : « *avoir une faim de loup* » = être vorace, insatiable ; « *se fourrer, se jeter, se précipiter dans la gueule du loup* » = courir à sa propre perte ; « *enfermer, laisser entrer le loup dans la bergerie* » = agir de façon inconsciente ; « *avoir vu le loup* » = pour une jeune fille, ne plus être vierge ; « *la faim fait sortir le loup du bois* » = se découvrir, se dévoiler ; « *l'homme est un loup pour l'homme* » = être féroce, impitoyable ; « *avancer à pas de loup* » = à pas feutré, sournoisement. Autant d'expressions qui pourraient faire l'objet d'un commentaire dans le cadre du *Petit Chaperon rouge*.

- *Ma mère-grand, que vous avez de grands bras !*
- *C'est pour mieux t'embrasser, ma fille.*
- *Ma mère-grand, que vous avez de grandes jambes !*
- *C'est pour mieux courir, mon enfant.*
- *Ma mère-grand, que vous avez de grandes oreilles !*
- *C'est pour mieux écouter, mon enfant.*
- *Ma mère-grand, que vous avez de grands yeux !*
- *C'est pour mieux voir, mon enfant.*
- *Ma mère-grand, que vous avez de grandes dents !*
- *C'est pour te manger.*

Comme on le voit, le lien entre la dominante et la sous-dominante thématique est très étroit.

b) Les 24 éléments de la « sous-dominante thématique »

Nous laissons au lecteur le soin d'identifier les vocables en fonction du déroulement de l'histoire. On repère aisément les verbes : *voir, courir, écouter, coucher, manger*, ils suivent la trajectoire dramatique. Auxquels s'ajoutent les flexions : *dit, envoie, apporte, tire, cherra, s'ouvrit*, qui tracent une ligne complémentaire de la trajectoire.

Bref, ce sont tous les vocables que l'on peut examiner à la loupe pour « redécouvrir » le texte, « analyser » le discours dans ses moindres détails et en « percevoir » toutes les finesses.

Mais, nous tenons à le souligner encore une fois, l'analyse doit mettre le texte au cœur de la composition pour en retirer tout le bénéfice analytique, car analyser (*αναλυειν* = défaire, décomposer) ne signifie pas « défaire pour défaire », pour séparer ou pour isoler, mais pour observer les liens et les agencements des éléments qui s'enchevêtrent pour former la trame du texte et du discours.

Aussi pour illustrer ce fait retiendrons-nous l'exemple des séquences relatives aux 23 emplois de la conjonction de coordination « *et* », qui sont automatiquement extraites par STABLEX en respectant ce que nous appelons « l'ordre chronologique » du texte :

Extraits de séquences concernant l'emploi de « et »

- *sa mère en était folle, et sa mère-grand plus folle encore.*
- *Un jour sa mère ayant cuit et fait des galettes, lui dit :*
- *« Va voir comme se porte ta mère-grand, car on m'a dit qu'elle était malade, porte-lui une galette et ce petit pot de beurre.*
- *« Je vais voir ma Mère-grand, et lui porter une galette avec un petit pot de beurre que ma Mère lui envoie.*
- *je m'y en vais par ce chemin ici, et toi par ce chemin-là, et nous verrons à qui plus tôt y sera.*
- *Le Loup se mit à courir de toute sa force par le chemin qui était le plus court, et la petite fille s'en alla par le chemin le plus long, s'amusant à cueillir des noisettes, à courir après des papillons, et à faire des bouquets des petites fleurs qu'elle rencontrait.*
- *« C'est votre fille, le Petit Chaperon rouge, dit le Loup, en contrefaisant sa voix,*

qui vous apporte une galette, et un petit pot de beurre que ma Mère vous envoie.
 – *Le Loup tira la chevillette, et la porte s'ouvrit.*
 – *Il se jeta sur la bonne femme, et la dévora en moins de rien;*
 – *Ensuite il ferma la porte, et s'alla coucher dans le lit de la mère-grand, en attendant le Petit Chaperon rouge, qui quelque temps après, vint heurter à la porte.*
 – *« C'est votre fille, le Petit Chaperon rouge, qui vous apporte une galette et un petit pot de beurre que ma Mère vous envoie.*
 – *Le Petit Chaperon rouge tira la chevillette, et la porte s'ouvrit.*
 – *« Mets la galette et le petit pot de beurre sur la huche, et viens te coucher avec moi.*
 – *Le Petit Chaperon rouge se déshabille, et va se mettre dans le lit, où elle fut bien étonnée de voir comment sa mère-grand était faite en son déshabillé;*
 – *Et, en disant ces mots, ce méchant Loup se jeta sur le Petit Chaperon rouge, et la mangea.*
 – *On voit ici que de jeunes enfants,*
Surtout de jeunes filles,
Belles, bien faites, et gentilles,
Font très mal d'écouter toute sorte de gens,
Et que ce n'est pas chose étrange,
S'il en est tant que le loup mange.
 – *Je dis le loup, car tous les loups*
Ne sont pas de la même sorte:
Il en est d'une humeur accorte,
Sans bruit, sans fiel et sans courroux,
Qui privés, complaisants et doux,
Suivent les jeunes Demoiselles
Jusque dans les maisons, jusque dans les ruelles

Nous laissons au lecteur le soin de lire attentivement ces séquences pour voir à quel point cette conjonction de coordination, qui est somme toute banale dans la typologie grammaticale, ne revêt pas toujours et systématiquement la même et unique valeur. On peut en retenir quelques-unes, par exemple :

1. une valeur d'opposition et aussi de renchérissement entre la mère et la grand-mère : « *sa mère en était folle, et sa mère-grand plus folle encore* » ;
2. une valeur d'abondance « *cuit et fait des galettes* » ;
3. une valeur d'opposition et de défiance : « *je m'y en vais par ce chemin ici, et toi par ce chemin-là, et nous verrons à qui plus tôt y sera.* »
4. une valeur de conclusion : « *Le Loup tira la chevillette, et la porte s'ouvrit.* »
5. une valeur hyperbolique : « *Belles, bien faites, et gentilles* » ; « *Sans bruit, sans fiel et sans courroux, / Qui privés, complaisants et doux* »

Chacun se fera une idée de ce qu'il faut faire pour mener à bien une analyse correcte, à la fois vérifiée et vérifiable, au vu des conclusions qui retiennent notre attention sur un mot « grammatical » aussi simple que la conjonction dite de coordination « *et* ».

C'est ainsi qu'on peut passer en revue les 79 vocables qui forment la trame du discours du *Petit Chaperon rouge*. En bas de liste, on trouve les vocables qui servent à agrémenter le discours, laquelle est d'ailleurs amplement enrichie par les vocables de fréquence 1, dont les

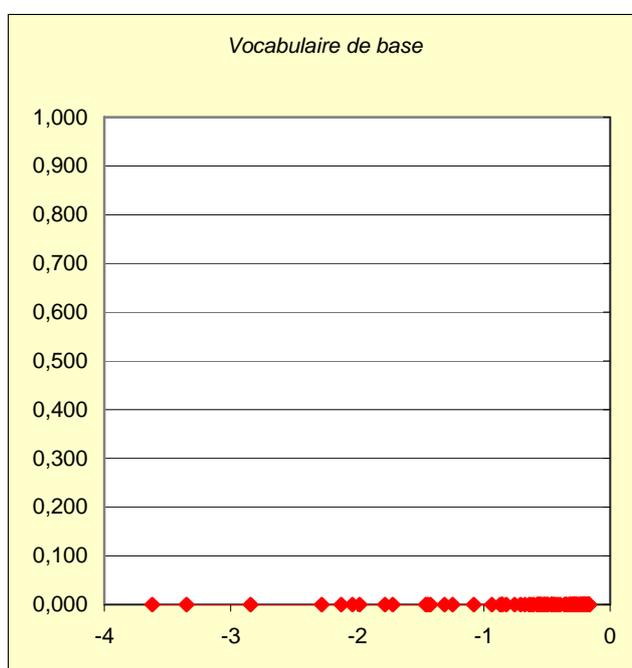
hapax, qui figurent dans la deuxième partie du vocabulaire, que nous qualifierons, par commodité, de « vocabulaire de base ». Ces éléments sont d'autant plus importants qu'ils sont affectés d'une densité hautement significative.

Bref, on voit bien que c'est toute l'étendue du vocabulaire qui est et doit être prise en considération. Comment pourrait-il en être autrement ? L'analyse ne doit rien rejeter, l'analyse ne peut rien rejeter, ne peut rien laisser de côté. Tous les éléments forment « texte et discours ».

4.2 Le vocabulaire de base

Paramètres de sélection : on sélectionne les valeurs négatives des résidus, $V_r < 0$, que l'on range suivant l'ordre croissant de dt

La branche négative de la parabole porte ainsi 192 vocables, dont 42 hapax qui font partie du vocabulaire propre de T2 :



Sur les 192 vocables qui figurent dans la partie négative, on dénombre 155 vocables de fréquence 1 dans T2 et 87 vocables avec une densité absolue hautement significative (ayant un $z > 2$, mis en gras sur la liste). De tous ces vocables, on remarque que ces sont les hapax qui sont dotés de la densité absolue la plus élevée ($z = 4,746$).

Mot	Occ	T2	<i>z</i>	<i>V_t</i>	<i>V_r</i>	<i>V_c</i>	<i>dt</i>	<i>V_r²</i>
peu	27	2	0,813	-0,202	-0,117	-0,165	0,002	0,014
entendit	9	1	1,020	-0,445	-0,250	-0,368	0,003	0,062
savait	9	1	1,020	-0,445	-0,250	-0,368	0,003	0,062
belles	10	1	0,901	-0,445	-0,266	-0,357	0,011	0,071
comment	10	1	0,901	-0,445	-0,266	-0,357	0,011	0,071
entrer	10	1	0,901	-0,445	-0,266	-0,357	0,011	0,071
moi	10	1	0,901	-0,445	-0,266	-0,357	0,011	0,071
peur	10	1	0,901	-0,445	-0,266	-0,357	0,011	0,071

première	10	1	0,901	-0,445	-0,266	-0,357	0,011	0,071
mettre	8	1	1,157	-0,445	-0,233	-0,380	0,017	0,054
quelques	8	1	1,157	-0,445	-0,233	-0,380	0,017	0,054
veux	8	1	1,157	-0,445	-0,233	-0,380	0,017	0,054
voyant	8	1	1,157	-0,445	-0,233	-0,380	0,017	0,054
ensuite	11	1	0,796	-0,445	-0,282	-0,346	0,025	0,080
force	11	1	0,796	-0,445	-0,282	-0,346	0,025	0,080
hélas	11	1	0,796	-0,445	-0,282	-0,346	0,025	0,080
jours	11	1	0,796	-0,445	-0,282	-0,346	0,025	0,080
longtemps	11	1	0,796	-0,445	-0,282	-0,346	0,025	0,080
ces	29	2	0,706	-0,202	-0,150	-0,142	0,030	0,023
eut	29	2	0,706	-0,202	-0,150	-0,142	0,030	0,023
d'abord	7	1	1,316	-0,445	-0,217	-0,391	0,031	0,047
faites	7	1	1,316	-0,445	-0,217	-0,391	0,031	0,047
loin	7	1	1,316	-0,445	-0,217	-0,391	0,031	0,047
très	7	1	1,316	-0,445	-0,217	-0,391	0,031	0,047
chez	12	1	0,701	-0,445	-0,299	-0,334	0,039	0,089
forêt	12	1	0,701	-0,445	-0,299	-0,334	0,039	0,089
gens	12	1	0,701	-0,445	-0,299	-0,334	0,039	0,089
moins	12	1	0,701	-0,445	-0,299	-0,334	0,039	0,089
bruit	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
disant	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
grosse	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
humeur	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
long	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
mange	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
mangé	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
partout	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
sous	6	1	1,508	-0,445	-0,201	-0,402	0,045	0,040
allait	13	1	0,615	-0,445	-0,315	-0,323	0,053	0,099
moralité	13	1	0,615	-0,445	-0,315	-0,323	0,053	0,099
vint	13	1	0,615	-0,445	-0,315	-0,323	0,053	0,099
yeux	13	1	0,615	-0,445	-0,315	-0,323	0,053	0,099
toute	31	2	0,608	-0,202	-0,183	-0,120	0,058	0,033
arriver	5	1	1,746	-0,445	-0,184	-0,414	0,059	0,034
envie	5	1	1,746	-0,445	-0,184	-0,414	0,059	0,034
voit	5	1	1,746	-0,445	-0,184	-0,414	0,059	0,034
chose	14	1	0,537	-0,445	-0,331	-0,312	0,067	0,110
appelait	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
étonnée	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
faite	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
maisons	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
mangea	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
mots	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
oui	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
partit	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
sera	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
su	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
t'	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
tôt	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
trouvait	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028
voyez	4	1	2,057	-0,445	-0,168	-0,425	0,073	0,028

bois	15	1	0,464	-0,445	-0,348	-0,300	0,081	0,121
demanda	15	1	0,464	-0,445	-0,348	-0,300	0,081	0,121
trois	15	1	0,464	-0,445	-0,348	-0,300	0,081	0,121
demeurait	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
dents	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
doux	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
embrasser	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
moulin	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
petites	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
porter	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
rencontra	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
toi	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
viens	3	1	2,497	-0,445	-0,152	-0,436	0,087	0,023
aller	20	2	1,275	-0,202	-0,003	-0,244	0,096	0,000
m'	20	2	1,275	-0,202	-0,003	-0,244	0,096	0,000
après	34	2	0,472	-0,202	-0,232	-0,086	0,100	0,054
bras	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
court	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
croyant	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
cueillir	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
demoiselles	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
dis	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
eh	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
fleurs	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
heurter	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
jambes	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
jolie	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
méchant	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
mets	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
osa	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
par-delà	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
passant	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
rencontrait	2	1	3,207	-0,445	-0,135	-0,447	0,102	0,018
accorte	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
adouçissant	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
amusant	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
arrêter	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
attendant	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
bouquets	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
bûcherons	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
cachant	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
chemin-là	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
compare	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
complaisants	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
contrefaisant	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
courroux	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
couverture	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
cuit	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
demeure-t-elle	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
déshabille	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
déshabillé	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
dévora	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014

doucereux	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
enrhumée	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
étrange	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
ferma	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
fiel	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
font	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
galettes	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
gentilles	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
heurte	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
huche	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
malade	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
noisettes	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
oh	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
oreilles	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
papillons	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
porte-lui	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
privés	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
ruelles	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
sait	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
seyait	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
suivent	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
surtout	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
ta	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
verrons	1	1	4,746	-0,445	-0,119	-0,459	0,116	0,014
filles	19	1	0,219	-0,445	-0,413	-0,255	0,137	0,170
mit	20	1	0,166	-0,445	-0,429	-0,244	0,151	0,184
a	38	2	0,310	-0,202	-0,297	-0,040	0,156	0,088
jour	21	1	0,116	-0,445	-0,445	-0,233	0,165	0,198
quelque	21	1	0,116	-0,445	-0,445	-0,233	0,165	0,198
fut	39	2	0,272	-0,202	-0,313	-0,029	0,170	0,098
eût	22	1	0,069	-0,445	-0,462	-0,221	0,179	0,213
là	22	1	0,069	-0,445	-0,462	-0,221	0,179	0,213
nous	22	1	0,069	-0,445	-0,462	-0,221	0,179	0,213
répondit	22	1	0,069	-0,445	-0,462	-0,221	0,179	0,213
aussitôt	24	1	-0,020	-0,445	-0,494	-0,199	0,207	0,244
pauvre	24	1	-0,020	-0,445	-0,494	-0,199	0,207	0,244
alla	42	2	0,164	-0,202	-0,362	0,005	0,213	0,131
tant	25	1	-0,062	-0,445	-0,511	-0,187	0,221	0,261
aussi	26	1	-0,102	-0,445	-0,527	-0,176	0,235	0,278
faire	44	2	0,097	-0,202	-0,395	0,027	0,241	0,156
autre	27	1	-0,141	-0,445	-0,543	-0,165	0,250	0,295
fois	27	1	-0,141	-0,445	-0,543	-0,165	0,250	0,295
ayant	28	1	-0,178	-0,445	-0,560	-0,153	0,264	0,313
temps	32	1	-0,315	-0,445	-0,625	-0,108	0,320	0,391
des	103	5	0,304	0,525	-0,080	0,695	0,326	0,006
tous	51	2	-0,116	-0,202	-0,509	0,107	0,339	0,259
rien	35	1	-0,408	-0,445	-0,674	-0,074	0,362	0,454
même	37	1	-0,467	-0,445	-0,707	-0,052	0,390	0,499
avec	55	2	-0,226	-0,202	-0,574	0,152	0,395	0,330
femme	55	2	-0,226	-0,202	-0,574	0,152	0,395	0,330
cette	40	1	-0,549	-0,445	-0,756	-0,018	0,432	0,571
où	59	2	-0,328	-0,202	-0,640	0,197	0,451	0,409

encore	44	1	-0,650	-0,445	-0,821	0,027	0,488	0,674
y	80	3	-0,222	0,040	-0,557	0,435	0,498	0,310
mais	81	3	-0,244	0,040	-0,573	0,446	0,513	0,328
enfants	46	1	-0,698	-0,445	-0,854	0,050	0,516	0,729
étaient	46	1	-0,698	-0,445	-0,854	0,050	0,516	0,729
fit	46	1	-0,698	-0,445	-0,854	0,050	0,516	0,729
fait	47	1	-0,721	-0,445	-0,870	0,061	0,530	0,757
pour	137	6	0,075	0,768	-0,209	1,079	0,556	0,044
comme	51	1	-0,810	-0,445	-0,935	0,107	0,587	0,875
une	158	7	0,112	1,011	-0,126	1,316	0,603	0,016
je	106	4	-0,243	0,283	-0,555	0,729	0,616	0,308
pas	109	4	-0,300	0,283	-0,604	0,762	0,658	0,365
ce	130	5	-0,228	0,525	-0,521	1,000	0,705	0,271
tout	82	1	-1,360	-0,445	-1,441	0,457	1,022	2,077
du	100	2	-1,115	-0,202	-1,309	0,661	1,027	1,714
on	138	4	-0,787	0,283	-1,077	1,090	1,065	1,161
son	109	2	-1,250	-0,202	-1,456	0,762	1,153	2,120
n'	133	3	-1,140	0,040	-1,422	1,034	1,243	2,021
si	99	1	-1,598	-0,445	-1,719	0,649	1,261	2,954
ne	181	4	-1,361	0,283	-1,779	1,577	1,669	3,166
avait	150	2	-1,771	-0,202	-2,125	1,226	1,729	4,517
à	331	11	-0,836	1,981	-1,246	3,273	2,042	1,552
les	249	6	-1,440	0,768	-2,037	2,345	2,129	4,151
d'	194	2	-2,223	-0,202	-2,844	1,724	2,347	8,086
elle	290	7	-1,550	1,011	-2,281	2,809	2,457	5,201
il	376	11	-1,273	1,981	-1,981	3,781	2,674	3,923
l'	225	2	-2,499	-0,202	-3,350	2,074	2,782	11,220
qu'	346	6	-2,320	0,768	-3,621	3,442	3,491	13,111

Les valeurs des densités montrent avec quelle précision et avec quel soin le conteur cisèle son texte, à la recherche du mot ou du terme précis ou approprié, comme le disent les 156 vocables de fréquence 1 (dont 42 hapax) sur un total de 271 vocables (58% des effectifs).

Loin de nous l'idée de vouloir présentement tout dire ou tout analyser. Procédons plutôt comme précédemment au relevé de quelques séquences à propos des vocables que l'on peut facilement identifier et replacer dans le contexte.

• Les hapax et les vocables de fréquence 1 mettent en relief le côté pictural de l'écriture, qui est comme une espèce de broderie dans l'écriture du texte :

- « *bûcherons* », « *forêt* », « *bois* » sont des vocables qui marquent l'embarras du Loup lorsque les circonstances ne lui sont pas favorables. Pour Perrault, c'est aussi l'occasion de « faire sortir le Loup du bois »¹⁰, c'est-à-dire de l'obliger à se démasquer.

En passant dans un bois elle rencontra compère le Loup, qui eut bien envie de la manger ; mais il n'osa, à cause de quelques bûcherons qui étaient dans la Forêt.

¹⁰ « *La faim fait sortir le loup du bois* », « *sortir du bois* », c'est se manifester. Or, notre Loup était bel et bien affamé.

- « papillons », « noisettes », bouquets », « fleurs », « cueillir », sont autant de vocables qui mettent en évidence le contraste entre la nonchalance de la petite fille qui flâne en chemin et l’empressement du Loup qui veut arriver le premier, pressé qu’il est d’arriver à ses fins :

Le Loup se mit à courir de toute sa force par le chemin qui était le plus court, et la petite fille s’en alla par le chemin le plus long, s’amusant à cueillir des noisettes, à courir après des papillons, et à faire des bouquets des petites fleurs qu’elle rencontrait.

- « bras », « jambes », « oreilles », « yeux », « dents » sont autant de vocables qui marquent l’étonnement de la petite fille devant « le déshabillé » de la grand-mère, à savoir du Loup qui s’est glissé dans les draps de la Mère-grand :
 - *Ma mère-grand, que vous avez de grands bras !*
 - *Ma mère-grand, que vous avez de grandes jambes !*
 - *Ma mère-grand, que vous avez de grandes oreilles !*
 - *Ma mère-grand, que vous avez de grands yeux !*
 - *Ma mère-grand, que vous avez de grandes dents !*
- « embrasser », « courir », « écouter », « voir », « manger » sont, à l’inverse, autant de vocables qui singularisent les réponses du Loup :
 - *C’est pour mieux t’embrasser, ma fille.*
 - *C’est pour mieux courir, mon enfant.*
 - *C’est pour mieux écouter, mon enfant.*
 - *C’est pour mieux voir, mon enfant.*
 - *C’est pour te manger.*

• Les vocables de « haute fréquence » sont là pour retenir l’attention du lecteur sur les faits marquants. C’est le cas, par exemple, des répétitions que l’on observe dans les citations précédentes : *c’est pour mieux... c’est pour mieux... c’est pour mieux...* qui accentuent (par gradation) l’empressement du Loup à « manger » la pauvre enfant, qu’il vint de mettre en confiance, à qui il donne du « *ma fille* », du « *mon enfant* » par trois fois, avant de se jeter sur elle, non sans avoir dévoré auparavant la grand-mère. C’est toute une stratégie discursive qui se dessine derrière ce style apparemment dépouillée, mais ô combien élaboré.

• La répétition de « *chemin* » au début du conte, assortie des adverbes « *ici* » et « *là* », met en évidence la ruse du Loup qui choisit le chemin le plus court après avoir « *oui* » les explications de la petite fille :

- Eh bien, dit le Loup, je veux l’aller voir aussi ; je m’y en vais par ce chemin ici, et toi par ce chemin-là, et nous verrons à qui plus tôt y sera.
Le Loup se mit à courir de toute sa force par le chemin qui était le plus court, et la petite fille s’en alla par le chemin le plus long.

Bref, aucun détail ne nous échappe. Il faut suivre le chemin tracé par le vocabulaire et entrer dans le texte pour découvrir toutes les subtilités du discours, et en apprécier la finesse par une « re-lecture » soignée et attentive.

4. La lemmatisation

La lemmatisation est un procédé technique qui consiste à rassembler les séquences d'après une racine ou un lemme, qui peut aller du phonème le plus simple aux expressions les plus complexes, en passant par les vocables ayant la même isotopie sémique ou sémantique.

4.1 La lemmatisation par la MACRO

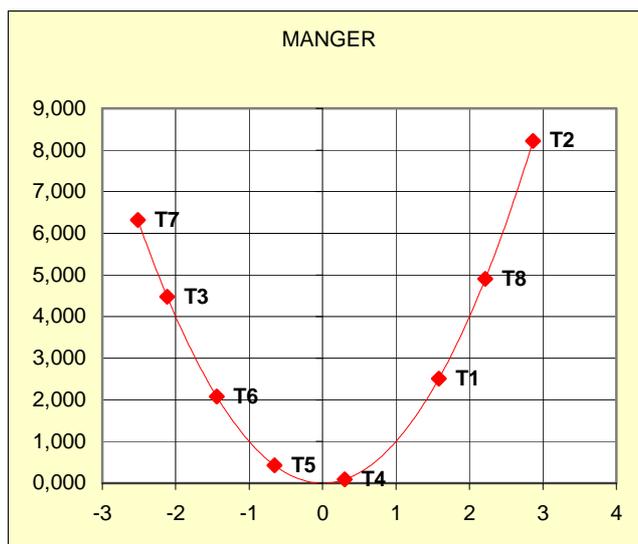
On choisit une « racine » dans la colonne des « mots » et on la fait apparaître dans la liste du vocabulaire extrait.

• Prenons, par exemple, le cas de « manger » que l'on mesure sur toute l'étendue du corpus :

Mot	Occ	T1	T2	T3	T4	T5	T6	T7	T8
manger	16	8	2		1	1	1		3
mangé	8	2	1		2				3
mange	4		1						3
mangea	4	1	1		1		1		
mangeaient	1								1
mangeât	1								1
mangèrent	1								1
mangés	1								1
MANGER	36	11	5	0	4	1	2	0	13
<i>densité</i>	0,241	1,584	2,867	-2,116	0,302	-0,654	-1,443	-2,515	2,215

On peut observer les différentes flexions en fonction des contes, assorties des densités appropriées. C'est ainsi que dans T2, *Le Petit Chaperon rouge*, (et aussi dans T8, *Le Petit Poucet*), le verbe *MANGER* revêt une importance capitale, malgré l'emploi limité à 5 occurrences, avec une densité de $z = 2,867$. La densité est en rapport direct avec l'intensité, et donc avec l'intention.

La « règle » qui figure dans la MACRO permet de mesurer la densité et d'en configurer l'image de distribution, suivant la parabole d'inertie :

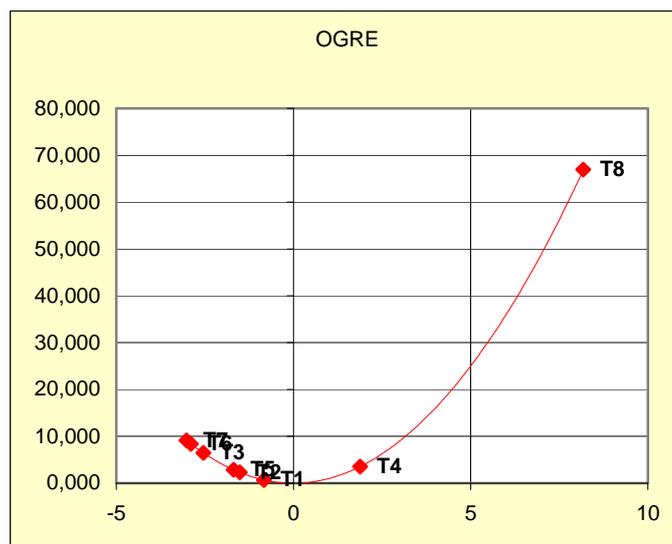


On suit parfaitement la trajectoire d'emploi du vocable « *manger* » dans le corpus tout entier, qui, de plus, est assorti des différentes flexions que l'on ne manquera pas d'étudier selon les perspectives d'analyse que l'on retient.

• Élargissons l'analyse aux 8 contes du corpus en prenant, par exemple, le cas de **OGRE**, un congénère du Loup :

Mot	Occ	T1	T2	T3	T4	T5	T6	T7	T8
ogre	41	1			9				31
ogresse	7	6							1
ogres	3	1							2
ogresses	1								1
OGRE	52	8	0	0	9	0	0	0	35
<i>densité</i>	-2,440	-0,831	-1,519	-2,543	1,878	-1,688	-2,893	-3,023	8,180

Le cas est immédiatement analysé et enregistré par le biais de l'image qui s'en dégage. La densité globale est négative : $z = -2,440$. Or, le personnage, sous ses différents aspects, n'apparaît que dans 3 contes, et n'occupe une place prépondérante que dans T8, *Le Petit Poucet* :



Il n'y a pas de limite à la recherche, laquelle est aisée, cohérente et riche, et en permanence vérifiée et vérifiable. De sorte qu'elle ne fait aucune place aux « divagations » ou aux « improvisations ». L'analyse doit être rigoureuse, vérifiable et vérifiée. Les conclusions doivent être constantes.

Certes, nous avons opéré sur des vocables notionnels, mais nous aurions pu aussi bien analyser l'emploi des possessifs, des démonstratifs, des adverbes, etc., ou encore opérer des regroupements plus vastes, suivant des familles grammaticales, thématiques, sémantiques...

4.2 La lemmatisation par STABLEX

Comme nous avons déjà vu les relevés d'emploi de la conjonction « *et* », nous pouvons passer au relevé des séquences concernant le lemme **MANGER** dans T2, *Le Petit Chaperon rouge* :

Mot : 'manger '

Extrait n° 1

En passant dans un bois elle rencontra compère le Loup, qui eut bien envie de la manger;

Extrait n° 2

car il y avait plus de trois jours qu'il n'avait mangé.

Extrait n° 3

C'est pour te manger.

Extrait n° 4

Et, en disant ces mots, ce méchant Loup se jeta sur le Petit Chaperon rouge, et la mangea.

Extrait n° 5

MORALITÉ

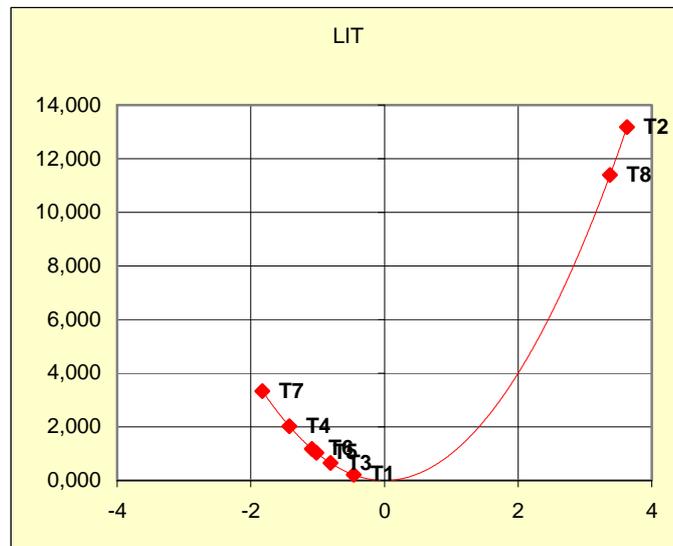
*On voit ici que de jeunes enfants,
Surtout de jeunes filles,
Belles, bien faites, et gentilles,
Font très mal d'écouter toute sorte de gens,
Et que ce n'est pas chose étrange,
S'il en est tant que le loup **mange**.*

Voilà une façon pratique de faire des dictionnaires, sachant que l'on peut élargir en permanence les champs de recherche, comme précédemment avec « manger » et « ogre ».

Les dictionnaires peuvent aussi être enrichis en permanence par des extraits provenant d'autres sources, si l'on veut mener à bien, par exemple, des études comparatives.

• La place du « **lit** » dans le corpus de T2. C'est un « espace hautement significatif » en fonction du double crime commis par le Loup :

Mot	Occ	T1	T2	T3	T4	T5	T6	T7	T8
lit	17	3	4						10
lits	2			1			1		
LIT	19	3	4	1	0	0	1	0	10
<i>densité</i>	0,387	-0,458	3,631	-0,806	-1,423	-1,020	-1,085	-1,827	3,376

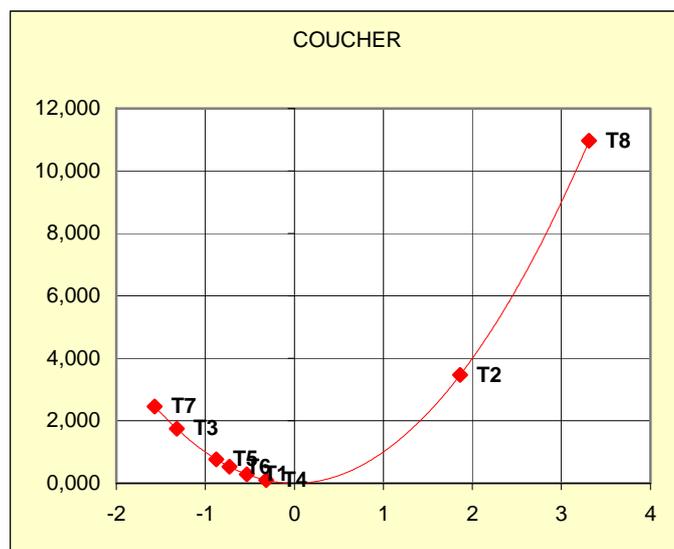


Dans le cas présent, il conviendra de vérifier si le « lit » a une place identique dans les 2 contes, dans T8 et dans T2 où le vocable ne laisse aucun doute quant à la densité d'emploi, et donc à l'intensité du discours et à l'intention du conteur.

• La place de « **coucher** » dans le corpus de T2 :

Mot	Occ	T1	T2	T3	T4	T5	T6	T7	T8
coucher	9		2						7
couché	3	2			1				
couchait	1						1		
couchés	1								1

COUCHER	14	2	2	0	1	0	1	0	8
densité	-0,168	-0,534	1,861	-1,320	-0,315	-0,876	-0,728	-1,568	3,312



Inutile d'aller plus loin dans les recherches et dans les commentaires. On voit combien l'analyse est précise, rigoureuse et incontestable. Tout est toujours vérifié et toujours vérifiable. Les nuages de points et les paraboles sont on ne peut lus parlants.

4.3 La discrimination par STABLEX

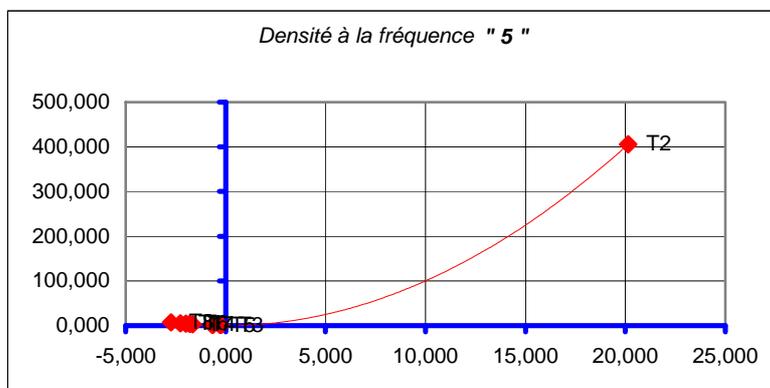
Discriminer revient à isoler un élément dans un ensemble complexe ou hétérogène. C'est une démarche opposée à celle de la lemmatisation. Il s'agit de différencier les emplois de termes homographes ou homonymes, comme entre le possessif « *son* » et le substantif « *son* », entre la forme verbale « *excellent* » et l'adjectif « *excellent* », ou encore, entre des vocables de même fréquence.

À titre d'exemple, nous pouvons regarder ce qui ce passe au niveau de la fréquence 5 :

Mot	Occ	T1	T2	T3	T4	T5	T6	T7	T8
arriver	5	1	1	1			1	1	
beurre	5		5						
courir	5		3		1				1
écouter	5		3						2
envie	5	2	1				1	1	
folle	5		2		1	1	1		
galette	5		5						
s'ouvrit	5		2	2				1	
pot	5		5						
village	5		3			1			1
voit	5		1	3				1	
voix	5	2	3						
Total	60	5	34	6	2	2	3	4	4

<i>densité</i>	8,800	-2,266	20,142	-0,260	-1,653	-0,648	-1,987	-1,796	-2,732
----------------	-------	--------	---------------	--------	--------	--------	--------	--------	--------

Inutile d'insister pour voir combien les vocables de fréquence 5 revêtent une importance capitale dans *Le Petit Chaperon rouge*, c'était annoncé dans la TDE (Table de contingence de Distribution des Écartés centrés réduits ou densités).



On peut alors isoler, par exemple, le qualificatif « **folle** » de forte densité de $z = 3,996$ dans T2. Le « *sens tropologique* » de ce qualificatif dit toute l'affection de la mère et de la grand-mère pour la chère enfant : « *sa mère en était folle, et sa mère-grand plus folle encore* ».

On peut également voir que le sens du substantif « **voix** » qui se rapporte au Loup, fait la part belle à la rhétorique par le biais de la prosopopée :

Mot : 'voix'

Extrait n° 1

« C'est votre fille, le Petit Chaperon rouge, dit le Loup, en contrefaisant sa **voix**, qui vous apporte une galette, et un petit pot de beurre que ma Mère vous envoie. »

Extrait n° 2

Le Petit Chaperon rouge, qui entendit la grosse **voix** du Loup, eut peur d'abord, mais croyant que sa mère-grand était enrhumée, répondit : « C'est votre fille, le Petit Chaperon rouge, qui vous apporte une galette, et un petit pot de beurre que ma Mère vous envoie. »

Extrait n° 3

Le Loup lui cria, en adoucissant un peu sa **voix** : « Tire la chevillette, la bobinette cherra. »

La recherche sur la page du Lexique de la MACRO peut déjà faciliter la tâche en donnant la localisation dans le corpus. Mais c'est avec STABLEX que l'on va opérer les relevés et les sélections.

Inutile d'alourdir cette présentation, on a compris de quoi il s'agissait. Seule la pratique peut ouvrir des horizons... et montrer que la recherche « scientifique » n'est pas un vain mot, même en matière d'analyse littéraire et discursive.

5. Conclusion

La conclusion sera rapide. Il n'est nullement besoin de grands commentaires pour voir quelles sont les performances de STABLEX dans la manipulation du texte et de l'hypertexte. Encore faut-il souligner que nous n'avons procédé que par petites touches pour ouvrir la voie à une analyse simple du *Petit Chaperon rouge*, le plus court des 8 contes de Perrault, et sans doute le plus facilement mémorisable.

Que dire alors des performances de STABLEX lorsqu'on aborde l'étude d'un large corpus où l'on analyse toutes les variables, dans toute leur richesse et leur complexité !

Le studieux qui refuserait de se contenter de redites incontrôlées, de lieux communs, d'approximations ou d'élucubrations, qui souhaiterait mener à bien une étude sérieuse où tous les éléments sont mesurés, contrôlés et vérifiés, et, par voie de conséquence, formuler des conclusions fondées sur des prémisses établies et vérifiables, bref, qui souhaiterait discuter et raisonner avec des arguments à l'appui, sans jamais perdre de vue les questionnements appropriés, n'hésitera pas à se lancer dans le bain et à se laisser guider dans les labyrinthes des matrices et des calculs matriciels, sans jamais perdre de vue qu'il s'agit de données qualitatives et non purement quantitatives : les densités et les intensités sont révélatrices des intentions.

Voyons, à titre d'exemple, le tableau comparatif des densités du corpus au vu des 10 valeurs les plus hautement significatives de T2, extraites de la TDR (Table des Valeurs Centrées Réduites) :

Rang	Densité	T1	T2	T3	T4	T5	T6	T7	T8	Fréq
78	3,291	-2,808	6,519	-2,449	-0,583	-0,778	9,036	-1,970	-3,676	15
41	2,983	-3,182	6,196	0,262	-0,086	-0,464	-0,468	1,231	-0,505	75
81	2,193	-1,637	5,741	-2,104	0,039	0,950	-2,162	-2,456	3,822	12
56	3,007	-0,745	4,880	0,859	-2,499	2,528	-0,267	-0,523	-1,226	39
70	1,230	2,129	4,423	-0,040	-0,214	-1,587	-1,015	-0,773	-1,693	23
88	1,045	1,467	4,144	0,513	-2,525	-0,900	0,984	-1,879	-0,759	5
79	2,148	-1,702	3,797	5,263	1,296	-2,006	-2,057	-0,689	-1,755	14
86	1,937	-0,276	3,504	-0,477	-2,178	2,528	-2,438	0,811	0,463	7
64	2,527	-0,921	2,823	-1,920	1,681	3,135	-2,191	0,003	-0,081	29
4	1,725	-1,201	2,783	-1,765	9,538	-2,625	-3,328	-1,481	-0,197	450

En bleu apparaissent les densités significatives négatives et en rouge les densités significatives positives. Bien entendu, nous laissons tout commentaire de côté. Il suffit de lire ces 10 lignes au gré des couleurs en suivant le fil des valeurs, pour se rendre compte de la précision que l'on apporte à l'analyse.

Plus encore, nous conseillons, pour toute initiation à la méthode, d'opérer sur des textes déjà connus, pour deux raisons évidentes : la première, c'est qu'on a la possibilité de vérifier les calculs, les données, le raisonnement et les conclusions ; la deuxième, c'est que, ce faisant, on entre de plain-pied dans la méthode, on la contrôle parfaitement et on se contrôle soi-même. Bref, on suit la trajectoire du discours en se laissant guider dans le labyrinthe du texte par le fil conducteur des densités lexicales, des intensités textuelles et des intentions

discursives. Que chacun se fasse sa propre idée à partir de sa propre expérience. Qui dit mieux ? Le jeu en vaut la chandelle.

La Statistique à la portée de tous

De la statistique pratique à la pratique de la statistique

10

Métrique R et ACP

Procédures graphiques

par
André CAMLONG
Christine CAMLONG-VIOT

Dans ce dixième chapitre, nous allons aborder les procédures de réalisation et d'adaptation des graphiques produits par la **MACRO** à la page de la *Métrique R*, suivant les principes de l'ACP (Analyse en Composantes Principales).

Nous invitons avant tout le lecteur à relire les descriptions figurant dans l'*Avertissement* inclus dans le pack de STABLEX.

*

* *

1. Les données de la *Métrique R*

Nous reprenons ici les données de la page de la *Métrique R* fournies dans le traitement des 8 contes en prose de Perrault.

	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
r	0,853	0,893	0,770	0,887	0,822	0,842	0,870	0,871	0,870
p	0,517	0,450	0,638	0,461	0,569	0,539	0,493	0,490	0,493
x	0,927	0,946	0,885	0,944	0,911	0,921	0,935	0,936	0,935
y	0,758	0,725	0,819	0,731	0,785	0,769	0,747	0,745	0,747
r'	0,706	0,786	0,540	0,774	0,645	0,685	0,740	0,743	0,740

ρ'	0,034	-0,099	0,276	-0,077	0,138	0,078	-0,013	-0,019	-0,014
r''	0,000	0,040	-0,083	0,034	-0,031	-0,011	0,017	0,018	0,017
ρ''	0,000	-0,066	0,121	-0,055	0,052	0,022	-0,023	-0,026	-0,024

Ces données sont automatiquement calculées à partir des valeurs des coefficients de corrélation r qui représentent la valeur du *cosinus* de l'angle formé par chaque variable avec l'axe d'origine OA, sur le cercle de centre O et de rayon 1.

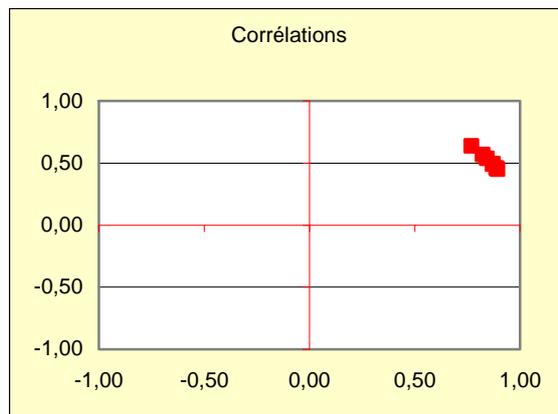
Nous allons détacher les graphiques se rapportant à chacune des 4 paires de données que nous commenterons chemin faisant.

2. Les cercles des corrélations

Rappel : $\cos = r$ et $\sin = \sqrt{1 - r^2}$, étant donné que $\cos^2 + \sin^2 = 1$ et que $r^2 + \rho^2 = 1$

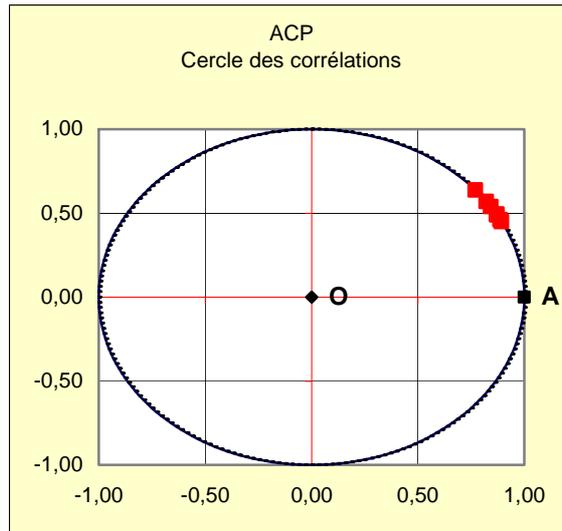
La première paire de données est composée du *cosinus* r et du *sinus* ρ , coordonnées de chaque point du nuage projeté sur le cercle des corrélations où il représente la variable correspondante.

Dans un premier temps, ce cercle est donné sous forme d'un « cercle carré » de centre O et de rayon 1, à savoir : un carré de côté compris entre -1 et +1 :



Comment transformer ce « cercle carré » en cercle des corrélations ?

Copier le cercle qui se trouve à la page *Estimation* de la Macro et le coller à côté du carré. Puis coller le carré sur le cercle et l'adapter.



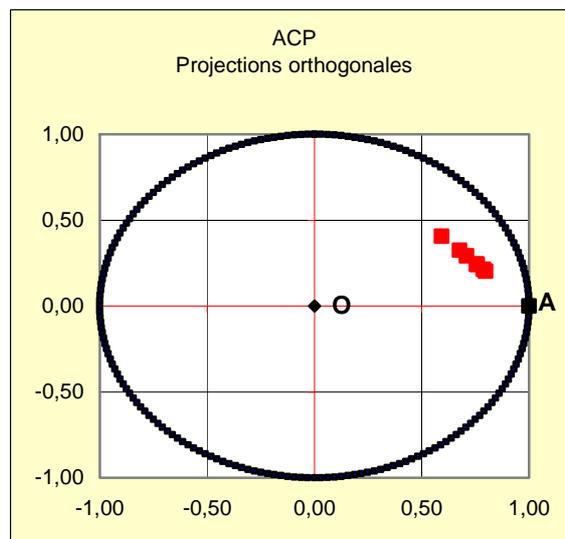
On voit que le nuage de points formé par les *cosinus* et les *sinus* des variables se projette sur le cercle.

Comment obtenir une projection orthogonale dans le cercle lui-même ?

Facile : il faut porter les carrés des cosinus r^2 et des sinus ρ^2 dans les deux lignes de la matrice en dessous des valeurs qui sont réservées à cet effet :

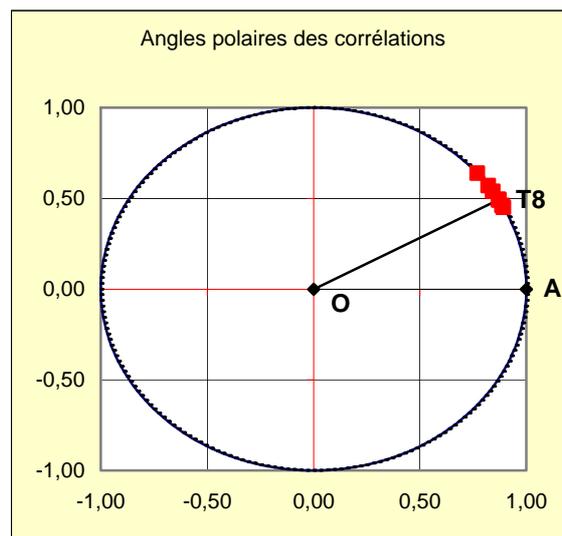
	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
r	0,853	0,893	0,770	0,887	0,822	0,842	0,870	0,871	0,870
ρ	0,517	0,450	0,638	0,461	0,569	0,539	0,493	0,490	0,493
r^2	0,729	0,797	0,593	0,787	0,676	0,710	0,757	0,759	0,757
ρ^2	0,271	0,203	0,407	0,213	0,324	0,290	0,243	0,241	0,243

Projeter ensuite les valeurs des coefficients de détermination (r^2 et ρ^2) à l'intérieur du cercle pour obtenir, sous forme d'ACP, le schéma des projections orthogonales :



Commentaires :

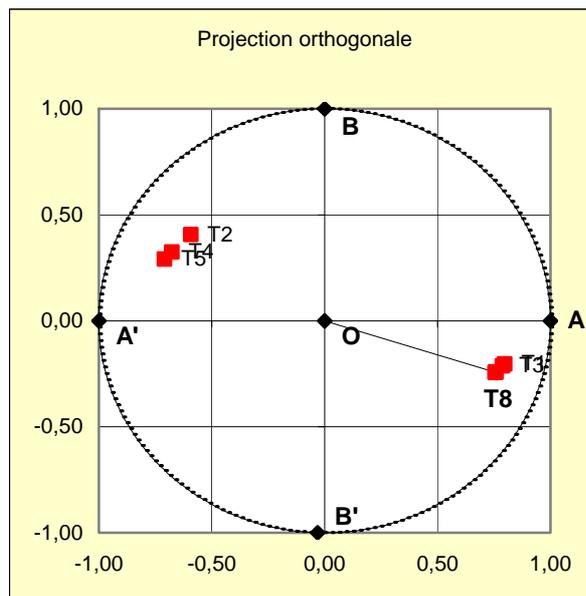
1. *Affichage des coordonnées.* Que ce soit sur le carré ou sur le cercle, les valeurs de chaque point sont affichées dès lors que le curseur les pointe.
2. *Calcul de l'angle.* Le calcul de l'angle formé par le point du nuage et l'axe des abscisses se fait avec les fonctions ACOS et DEGRES. Exemple : pour un angle moyen du nuage, de cosinus 0,853, d'arc cosinus 0,549, l'angle vaut $31^{\circ}26'$.
3. *Valeurs propres.* L'angle le plus faible est celui qui a le cosinus le plus élevé, comme celui de T1 qui, pour un cosinus de 0,893 forme un angle de $26^{\circ}46'$. Et, inversement, l'angle le plus grand est celui qui a le cosinus le plus faible, comme celui de T2, qui forme un angle de $39^{\circ}39'$ pour un cosinus de 0,770. Résultat : T1 est proche de l'origine A (puisque *cosinus* $\rightarrow 1$) sur le cercle, et, inversement, T2 s'en éloigne (*sinus* $\rightarrow 1$), trigonométrie oblique.
4. *Seuils de signification.* Lorsque les points du nuage se situent dans le carré le plus proche de l'origine A (où *cosinus* $r = 1$), la corrélation est fortement marquée. Inversement, lorsqu'ils s'en éloignent (où *sinus* $\rho \rightarrow +1$), les variables résistent à la corrélation. Le seuil de signification est posé pour un angle de 30° (où *cosinus* $r = \sqrt{3}/2 = 0,866^1$).
5. *Arcs de cercle et angles polaires.* Le point T8 dessine un arc de cercle AT₈ avec un angle polaire θ (AOT₈) de $29^{\circ}32'$ sur le cercle des corrélations :



6. *Vecteurs propres.* On évitera de confondre le *cosinus* de l'angle qui est sur le cercle des corrélations (*cos* $r =$ coefficient de corrélation) et le *cosinus carré* de la projection orthogonale du point dans le cercle (*cosinus carré* $r^2 =$ coefficient de détermination). Le *cosinus*, c'est la longueur de l'arc de cercle, alors que le *cosinus carré* (r^2) est la longueur du côté adjacent à l'angle droit du triangle rectangle dont

¹ Voir supra Cap. 2, page 27. Néanmoins, il ne s'agit là que de tendances générales ou globales, qui seront affinées par l'AFD, étant donné que la *corrélation est intransitive*, que *les variables sont corrélées par paires*. Toute généralisation du modèle est de ce fait arbitraire. Le but de l'ACP est de « faire voir », pour « faire connaître » et « faire reconnaître » par le biais d'une « image phénoménale » de la matrice de corrélation. En outre, soulignons-le une fois de plus, les données de corrélation ont une *dimension géométrique*, alors que les densités ont une *valeur algébrique dans une fonction d'intégration, de description et de comparaison*. Voilà pourquoi nous attachons la plus grande importance au schéma fourni par les vecteurs isotropes, qui reflètent amplement la « hiérarchie naturelle » des variables d'un corpus. Qu'est-ce que la géométrie euclidienne, sinon « l'art de raisonner juste sur des figures fausses », dit-on ? C'est ce que les philosophes appellent « la certitude de droit ». S'il en est ainsi, gageons que l'ACP a atteint son objectif.

l'hypoténuse $OT_8 = 1$, en vertu du théorème de Pythagore : $\cos^2 + \sin^2 = 1$. Le vecteur OT_8 est *vecteur propre* (= hypoténuse de la projection orthogonale) :



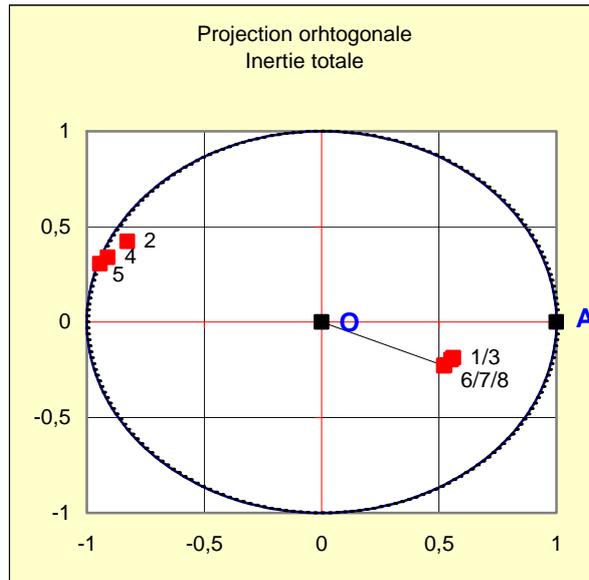
7. *Configuration orthogonale.* La longueur de chaque *vecteur propre* $OT_x = 1$, longueur de l'hypoténuse du triangle rectangle formé par chaque point du nuage se projetant orthogonalement dans la même sphère. Le graphique donne une coupe plane de la projection autour du centre de gravité $O(0;0)$.
8. *Fonctions circulaires.* Rappels : 1) Les fonctions circulaires sont continues pour toutes les valeurs de x : la dérivée de $y = \sinus x$ c'est $y' = \cos x$, et de $y = \cos x$, c'est $y' = -\sinus x$. 2) Dans un triangle, les côtés sont proportionnels aux *sinus* des angles opposés : $\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$.
9. *Vecteurs isotropes.* Les relations trigonométriques du triangle rectangle font qu'on retiendra les *sinus* pour mesurer les *vecteurs isotropes*. (Voir *infra* § 5.4)

Néanmoins, compte tenu des remarques précédentes, et sachant que le coefficient de détermination r^2 joue un rôle capital en statistique puisqu'il donne : 1) d'une part, en fonction du *cosinus carré* r^2 , le pourcentage de la variance contrôlée, et 2) d'autre part, en fonction du *sinus carré* ρ^2 , le pourcentage de la variance non contrôlée (des résidus), il est important de *polariser* le système bilinéaire sous la forme quadratique centrée réduite autour du centre de gravité $O(0;0)$. Alors l'inertie est totale. Les valeurs sont en tout point semblables aux valeurs centrées réduites des densités (de Laplace-Gaus).

D'où la transformation du tableau précédent en tenant compte du signe d'inertie des cosinus et des sinus :

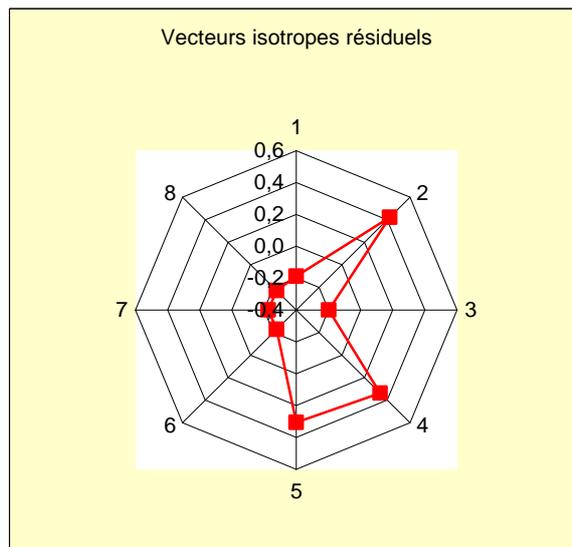
	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
r	0,853	0,893	0,770	0,887	0,822	0,842	0,870	0,871	0,870
ρ	0,517	0,450	0,638	0,461	0,569	0,539	0,493	0,490	0,493
r^2	0,235	0,797	-0,593	0,787	-0,676	-0,710	0,757	0,759	0,757
ρ^2	-0,015	-0,203	0,407	-0,213	0,324	0,290	-0,243	-0,241	-0,243
r^2 centré	0,000	0,562	-0,828	0,552	-0,911	-0,944	0,522	0,525	0,522
ρ^2 centré	0,000	-0,188	0,422	-0,198	0,339	0,306	-0,228	-0,225	-0,228

D'où la projection orthogonale d'inertie maximale autour du centre $O_{(0;0)}$:



L'avantage de la transformation du système bilinéaire en cercle polaire est de pouvoir introduire d'autres variables dans le schéma ou de pouvoir comparer plusieurs schémas entre eux, et donc plusieurs matrices entre elles.

Il en est de même du cercle polaire des vecteurs isotropes qui donnent la mesure comparée des forces résiduelles en fonction de la « longueur » des *sinus centrés réduits* ρ^2 :



Comme la distance des points au centre de gravité, épicode du système, est égale à la moyenne des carrés des distances de tous les individus, en tant que variable centrée réduite, elle est d'un intérêt immédiat en statistique descriptive.

Rappelons que le but de l'ACP (Analyse en Composantes Principales) est d'explorer au mieux la structure de la matrice, et donc des données elles-mêmes².

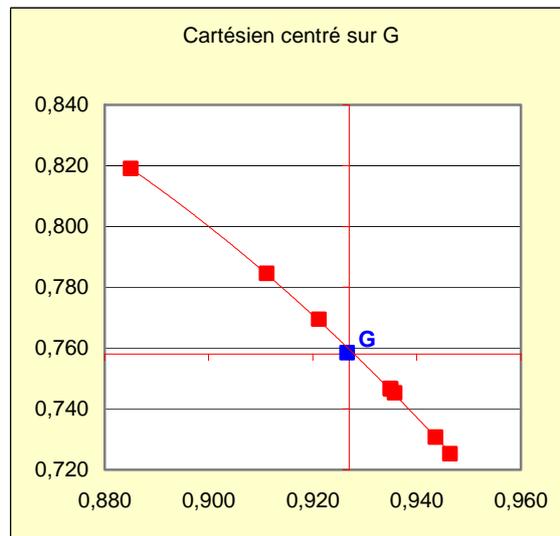
3. Les coordonnées cartésiennes x et y

$$\text{Rappel : } x = \frac{r+1}{2} \text{ et } y = \frac{\rho+1}{2}$$

Prenons les données correspondantes :

	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
x	0,927	0,946	0,885	0,944	0,911	0,921	0,935	0,936	0,935
y	0,758	0,725	0,819	0,731	0,785	0,769	0,747	0,745	0,747

Sélectionner les valeurs des variables, tracer le graphique en nuage de points et placer le centre de gravité G (0,927 ; 0,758), barycentre d'abscisse 0,927 et d'ordonnée 0,758 :



Projeté à l'échelle normale, le nuage de points est écrasé. Mais, grossi autour du centre de gravité, il donne une image visible et lisible du schéma des variables sur l'arc de cercle.

C'est sur cette « configuration unique » que l'ACP prend en charge les transformations des équations linéaires pour présenter un ensemble de figures qui font miroiter le même schéma sous des angles ou des biais différents.

4. Projections hiérarchiques et orthogonales avec r' et ρ'

$$\text{Rappel : } r' = 2r - 1 \text{ et } \rho' = 2\rho - 1$$

² Ces 2 schémas des projections orthogonales ne diffèrent que par la qualité du centre de gravité qui donne la mesure du vecteur propre : d'un côté, O est un centre d'inertie, et, de l'autre, G est un barycentre. Mais, dans les 2 cas, la comparaison est immédiate : elle est à la mesure de la distance qu'il y a entre les individus, une distance sans dimension. Comme la hiérarchie entre les individus est fixe, l'image est stable malgré les apparences.

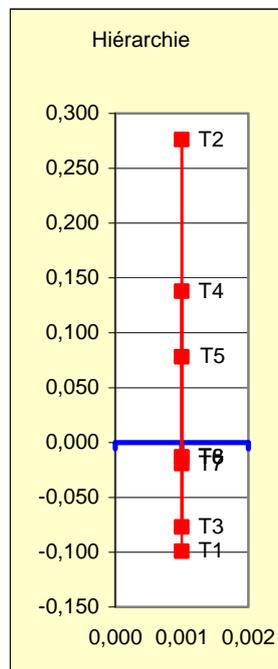
4.1 Projection du sinus l'axe correspondant (vertical)

Comment faire ? Donner une valeur constante au *cosinus* dans la ligne supérieure réservée à cet effet, laquelle *Constante* est l'abscisse du couple de coordonnées, l'ordonnée étant ρ' .

Prenons les données correspondantes :

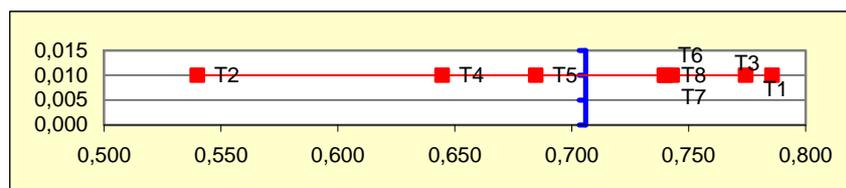
	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
C^{te}		0,010	0,010	0,010	0,010	0,010	0,010	0,010	0,010
r'	0,927	0,946	0,885	0,944	0,911	0,921	0,935	0,936	0,935
ρ'	0,758	0,725	0,819	0,731	0,785	0,769	0,747	0,745	0,747

Le *sinus* ρ' se projette sur l'axe vertical du graphique qui prend la forme d'un cadran de thermomètre. On peut alors lire sur ce cadran gradué la « hiérarchie naturelle » des points qui est à l'image des variations de température :



Travailler ensuite le graphique à l'aide des poignées pour adapter tailles et dimension.

NB : On pourrait de la même façon la montrer à partir du *cosinus* r' , elle serait identique à la précédente, mais l'image serait verticale avec une échelle différente (trigonométrie oblique) :



Quoi qu'il en soit, ces 2 schémas donnent la « hiérarchie naturelle » des points.

Le but de l'ACP est de « faire voir », pour « faire connaître » et « faire reconnaître », afin d'orienter et de faciliter le raisonnement. C'est tout l'art de la géométrie descriptive.

4.2 Projection orthogonale

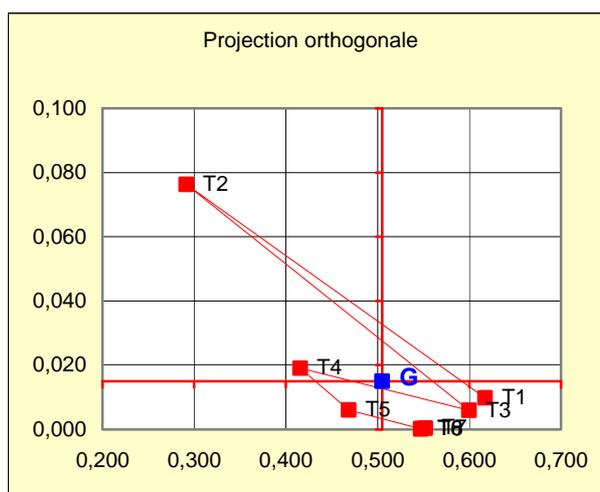
La projection orthogonale fait appel au carré des cosinus r' et sinus ρ' :

Pour ce faire, on calcule les carrés des cosinus et des sinus dans les lignes réservées à cet effet, à savoir : r'^2 , le carré de r' dans la ligne supérieure, et ρ'^2 , le carré de ρ' dans la ligne inférieure.

Prenons les données correspondantes :

	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
r'^2	0,505	0,617	0,292	0,600	0,416	0,469	0,547	0,552	0,548
r'	0,927	0,946	0,885	0,944	0,911	0,921	0,935	0,936	0,935
ρ'	0,758	0,725	0,819	0,731	0,785	0,769	0,747	0,745	0,747
ρ'^2	0,015	0,010	0,076	0,006	0,019	0,006	0,000	0,000	0,000

Le graphique reproduit l'image centrée de la projection orthogonale :



Évidemment, cette image n'est qu'une image plane. Elle serait sans doute plus suggestive et plus conforme à la réalité des faits si elle était projetée dans une sphère de rayon 1. Néanmoins, elle éclaire sur la géométrie spatiale des points projetés sur une surface plane et distribués autour du centre de gravité G $(0,505 ; 0,015)$, (barycentre des polynômes).

Dans ce nuage de points, on voit non seulement le point T2 se détacher et occuper une place particulière dans l'ensemble, mais on voit aussi les autres points se distribuer dans des espaces conventionnels, plus ou moins distants ou rapprochés.

De fait, cette image reflète les qualités inhérentes à la matrice des corrélations, qualités établies par l'algèbre et « affichées » par les schémas (géométrie oblige)³.

³ Voir in *Cahiers II, La Pléiade*, p. 783 comment Paul Valérie voit dans la géométrie *un instrument de pensée*, au sens étymologique du latin *pesare*, un instrument qui permet à l'esprit de *voir*, au sens étymologique du verbe grec $\epsilon\iota\delta\omega$: « L'important et le beau de la géométrie c'est (par sa *pureté*) qu'elle est un instrument de pensée – un mode de traitement – une manière de *voir* et de prolonger et non un objet étranger. Tout ce qui permet de bien

5. Projections obliques et rotations avec inertie totale en r'' et ρ''

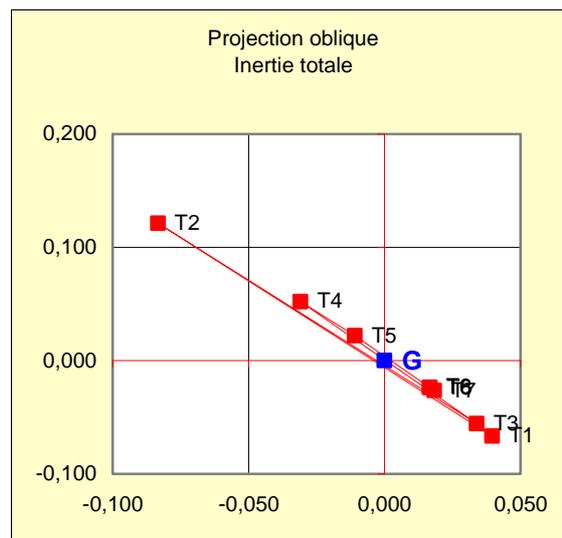
Rappel : $r'' = r - \bar{r}$ et $\rho'' = \rho - \bar{\rho}$

Prenons les données correspondantes :

	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
r''	0,000	0,040	-0,083	0,034	-0,031	-0,011	0,017	0,018	0,017
ρ''	0,000	-0,066	0,121	-0,055	0,052	0,022	-0,023	-0,026	-0,024

5.1 Projection oblique avec inertie totale

La projection oblique a une qualité fondamentale qui tient à « l'inertie totale » du schéma. Les points du nuage se distribuent autour du centre de gravité $G_{(0;0)}$:



Les points ne se situent pas, à strictement parler, sur la même ligne. Les uns sont décalés vers le haut et les autres vers le bas, en fonction des qualités fondamentales de la distribution, comme le montrent les figures circulaire, cartésienne ou orthogonale.

C'est toujours la même image, comme une espèce d'image bloquée, vue sous des angles différents, que l'on fait tourner pour mieux en observer les caractéristiques.

discerner et de fixer des opérations de l'esprit est de nature géométrique. Et toutes les définitions géométriques vraies sont des constructions ou opérations. Nous ne pouvons rien de plus. » Cette définition convient à merveille à l'épure de l'architecte, et donc à l'ACP. Alors qu'on n'a que des lignes et des cotes sur une feuille, l'esprit entraîné perce le secret des murailles tracées, il en perçoit les contours, les agencements, les lignes... la structure, l'architecture et la géométrie.

5.2 Rotation des projections obliques

On a beau faire tourner les schémas autour du centre de gravité, on voit que l'image est fixe, que l'inertie est totale.

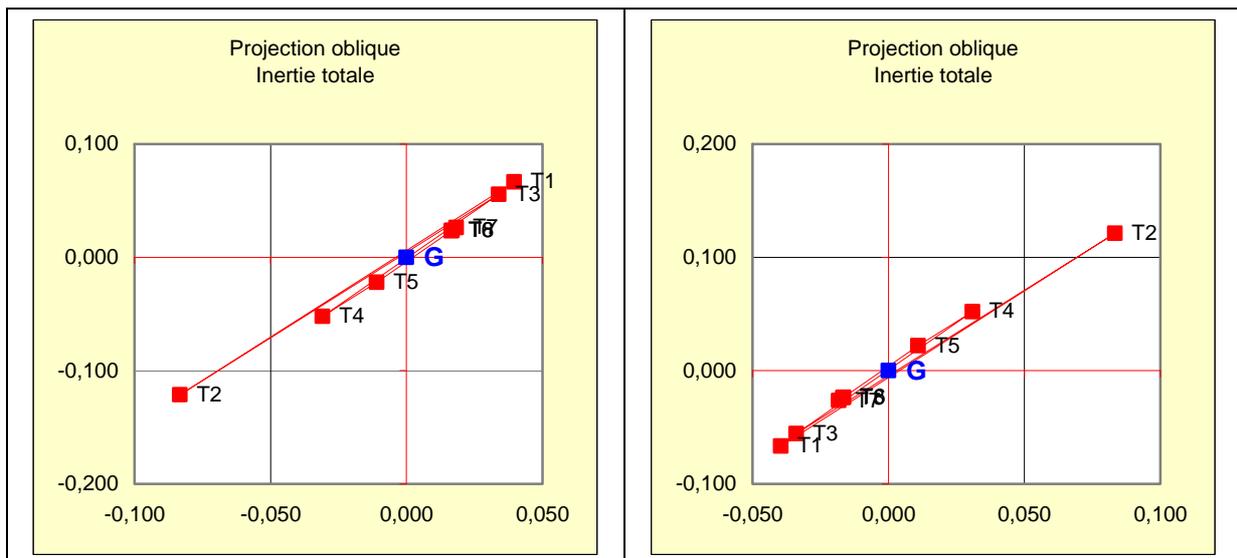
Comment la faire tourner de 90° , par exemple ? Il suffit de multiplier par -1 les valeurs du *cosinus* r'' et du *sinus* ρ'' et d'apparier les coordonnées pour que l'image se projette dans les quadrants du cercle selon les 4 schémas possibles, en fonction des 4 combinaisons.

On peut aussi la faire tourner « par petites rotations », comme les aiguilles sur le cadran d'une montre, en multipliant les coordonnées par un décimal compris entre -1 et $+1$.

Prenons les données multipliées par -1 qui permettent de former 4 paires :

	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
$(-1),r''$	0,000	-0,040	0,083	-0,034	0,031	0,011	-0,017	-0,018	-0,017
r''	0,000	0,040	-0,083	0,034	-0,031	-0,011	0,017	0,018	0,017
ρ''	0,000	-0,066	0,121	-0,055	0,052	0,022	-0,023	-0,026	-0,024
$(-1),\rho''$	0,000	0,066	-0,121	0,055	-0,052	-0,022	0,023	0,026	0,024

Prenons 2 paires de combinaisons pour voir 2 images inversées :



Que voit-on ? Deux figures identiques. *Des schémas fixes* pour une « hiérarchie fixe ».

Tout l'art de l'ACP est de mettre la puissance géométrique au service de la puissance descriptive, afin de *fixer l'idée*⁴ même des relations ou des liaisons qualitatives naturelles qu'il y a entre les variables d'un corpus.

5.3 Projections orthogonales d'inertie totale

Comment passer de la projection du nuage de points autour du centre de gravité $G(0; 0)$ à la projection orthogonale de la matrice dans le cercle des corrélations ?

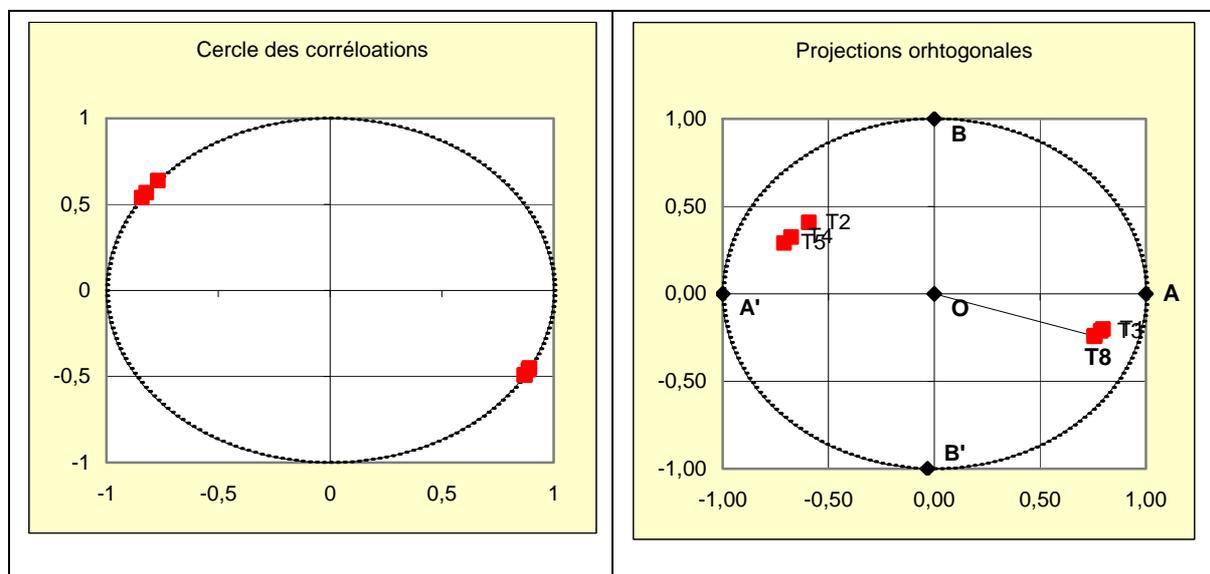
⁴ Idée vient du grec $\epsilon\iota\delta\omega$, qui signifie « voir ». L'image donne à voir des relations qualitatives qui vont au-delà des simples corrélations quantitatives. *Voir et savoir pour connaître et reconnaître*, c'est tout l'art et toute la puissance que la géométrie met au service de la pensée, pour mesurer et raisonner, juger et critiquer.

Comme pour toute projection orthogonale, il faut reprendre *le carré des cosinus et des sinus* (valeurs propres des variables) et les affecter du signe algébrique de l'inertie totale.

Reprenons le tableau des valeurs correspondantes :

	moyenne	T1	T2	T3	T4	T5	T6	T7	T8
r''	0,000	0,040	-0,083	0,034	-0,031	-0,011	0,017	0,018	0,017
ρ''	0,000	-0,066	0,121	-0,055	0,052	0,022	-0,023	-0,026	-0,024
r	0,245	0,893	-0,770	0,887	-0,822	-0,842	0,870	0,871	0,870
ρ	-0,080	-0,450	0,638	-0,461	0,569	0,539	-0,493	-0,490	-0,493
r^2	0,235	0,797	-0,593	0,787	-0,676	-0,710	0,757	0,759	0,757
ρ^2	-0,015	-0,203	0,407	-0,213	0,324	0,290	-0,243	-0,241	-0,243

En mettant côte à côte le cercle des corrélations et le cercle des projections orthogonales, on perçoit mieux le *positionnement des points* :



Les images proposées sont représentatives tant de *la qualité de la corrélation* que de *la distance entre les points*.

Voilà comment le parcours analytique proposé par la *métrique R* est un parcours géométrique qui vise à mettre en évidence les qualités fondamentales des corrélations, dans le seul but d'ouvrir la voie à l'*AFD* (Analyse Factorielle des Données), une analyse qualitative par excellence.

5.4 Inertie totale des vecteurs isotropes

L'*inertie totale*, qui concerne les vecteurs isotropes, fournit des renseignements utiles sur *la résistance d'intégration inhérente aux variables*.

En effet, la longueur des vecteurs isotropes résume à merveille la finalité de l'ACP. L'image montre que les qualités inhérentes à la matrice des corrélations sont proportionnelles à la taille des vecteurs isotropes⁵.

On privilégie, trigonométrie oblige, la mesure des *résistances* à la force des *sinus*, qui reflètent les qualités essentielles des composantes résiduelles du corpus :

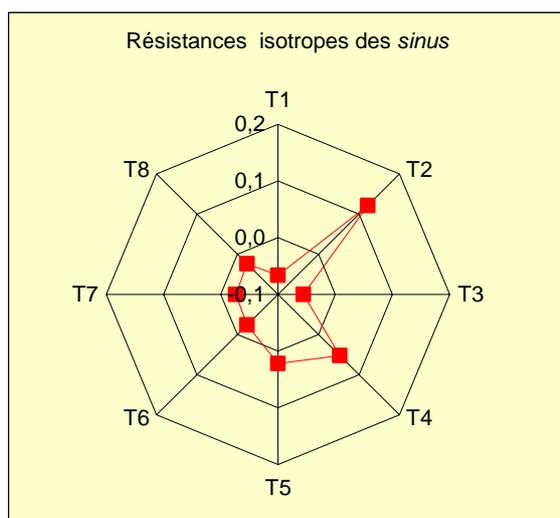
- 1) La fonction $y = \sinus x$ est une fonction affine dont la régularité et la continuité se retrouvent dans la *sinusoïde*
- 2) En outre, la valeur complémentaire du *cosinus* donne la pente en fonction de la dérivée $y' = \cos x$.

Le tout éclaire la partie et la partie éclaire le tout⁶.

Comparons les deux graphiques au vu des valeurs trigonométriques complémentaires : *sinus* et *cosinus*.

5.4.1 Les résistances isotropes des *sinus* (forces résiduelles)

La mesure des *vecteurs isotropes* est proportionnelle à la *résistance d'intégration* des *sinus*. La force de résistance est une force centrifuge, proportionnelle à la longueur d'onde du vecteur qui a son origine dans l'épicentre du graphique :



⁵ Bien qu'étant de pures figures géométriques (des épures), les schémas sont révélateurs des qualités intrinsèques des variables qui forment une matrice, dont la matière première n'est point le nombre qui prime, mais l'être qui le sous-tend. Malheur à celui qui ne voit dans les mathématiques que des nombres inertes, au lieu des rapports qualitatifs établis entre les êtres ! La longueur des vecteurs isotropes en est la parfaite illustration.

⁶ Dire avec Platon que le tout éclaire la partie et que la partie éclaire le tout, c'est le principe même de l'analyse. Conformément à l'étymologie grecque du terme αναλυω, *analyser*, c'est séparer pour voir comment c'est corrélé, dans quelles proportions, quel ne est l'équilibre fondamental. Nous pourrions prendre l'exemple de la montre comme métaphore explicative. La montre n'est pas un ensemble de pièces jetées dans un chapeau, mais un ensemble où les pièces s'articulent pour former un tout, en fonction de la finalité qui lui est dévolue par le créateur. Allons plus loin. Que dit-elle ou que donne-t-elle la montre ? Hélas ! elle ne dit rien, elle ne donne rien non plous, si ce n'est le centre de gravité de l'utilisateur, un centre trigonométrique qui lui permet de se situer dans le temps et dans l'espace, d'un point de vue mental, psychologique et physique, et donc de s'orienter. Prétendre le contraire, ce serait « perdre le nord », « perdre la boussole » ou « perdre la tête ». Le centre trigonométrique de la montre, c'est le fameux *ego, hic et nunc*, le « je, ici et maintenant ». Bref, la montre, c'est comme la boussole qui dirige les pas, qui trace le chemin, qui donne la mesure intérieure et extérieure, qui oriente...

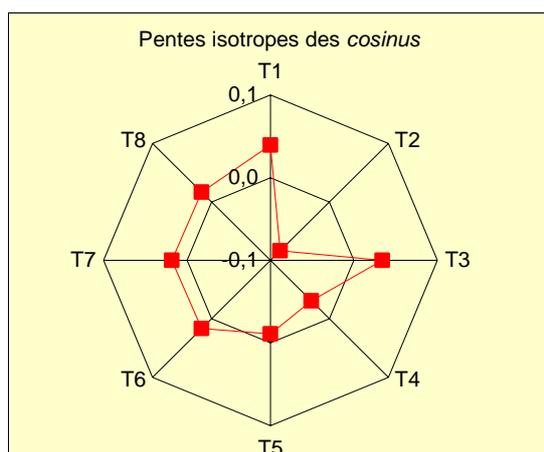
Le vecteur qui offre la plus grande résistance, c'est le vecteur le plus long, celui du point qui s'éloigne le plus de l'épicentre (point de convergence). Dans le graphique ci-dessus, c'est T2, *le plus grand des vecteurs isotropes*, qui montre *la plus grande distance ou la plus forte résistance*⁷.

5.4.2 Les pentes isotropes des *cosinus* (forces contrôlées)

Tel un paradoxe, les vecteurs isotropes en fonction des *cosinus* donnent une image inverse de la précédente, en vertu des relations trigonométriques entre *cosinus* r et *sinus* ρ . Cette image est représentative de la *pente sinusoidale* : $y' = \cos x$ est la dérivée de $y = \sin x$. Alors que $y' = -\sin x$ est la dérivée de la fonction $y = \cos x$.

Rappelons que les valeurs trigonométriques usuelles des sinus et des cosinus sont à la fois inversées et complémentaires.

C'est T2, le vecteur le plus faiblement corrélé (*cosinus* r le plus faible), qui est le plus résistant (*sinus* r'' le plus fort), qui a la pente la plus faible (*cosinus* r le plus faible), et qui est le plus proche de l'épicentre (*cosinus* r'' le plus faible) :



Rappel des coordonnées de T2 (à considérer dans la matrice) :

- coordonnées naturelles : *cosinus* $r = 0,770$; *sinus* $\rho = 0,638$
- coordonnées en inertie totale : *cosinus* $r'' = -0,083$; *sinus* $\rho'' = 0,121$.

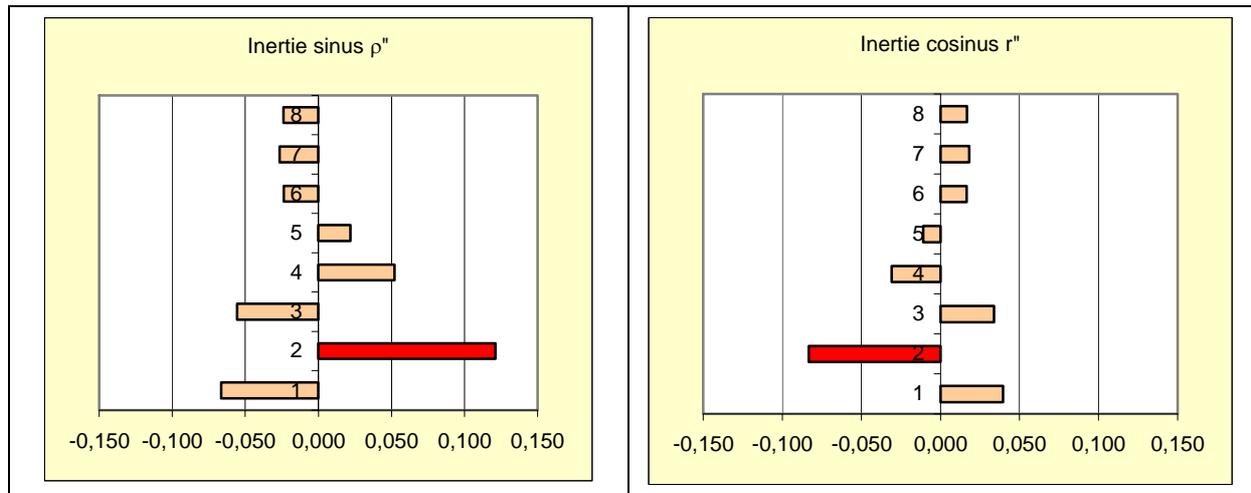
Comme on le voit, il est inutile de multiplier les graphiques, *ils disent tous la même chose*. Ils révèlent les qualités inhérentes à la matrice des corrélations, et donc aux composantes du corpus.

5.4.3 Symétrie des schémas en fonction du *cosinus* et du *sinus*

Comme les *cosinus* et les *sinus* sont complémentaires, on pourrait croire que les schémas sont symétriques, toutes proportions gardées. Mais, c'est une erreur grossière qui consiste à faire fi des échelles.

⁷ Comparer ce graphique au graphique des valeurs centrées réduites § 2 à la page 218. Bien que s'agissant de centre polaire O dans les deux cas, la différence est fondamentale : dans le premier cas, il s'agit du centre polaire des *coefficients de détermination*, et, dans le second cas, du centre polaire des *coefficients de corrélation*.

Pour éviter cette erreur, il suffit de mettre côte à côte les diagrammes des *sinus* ρ'' et des *cosinus* r'' et d'observer les échelles pour s'en persuader :



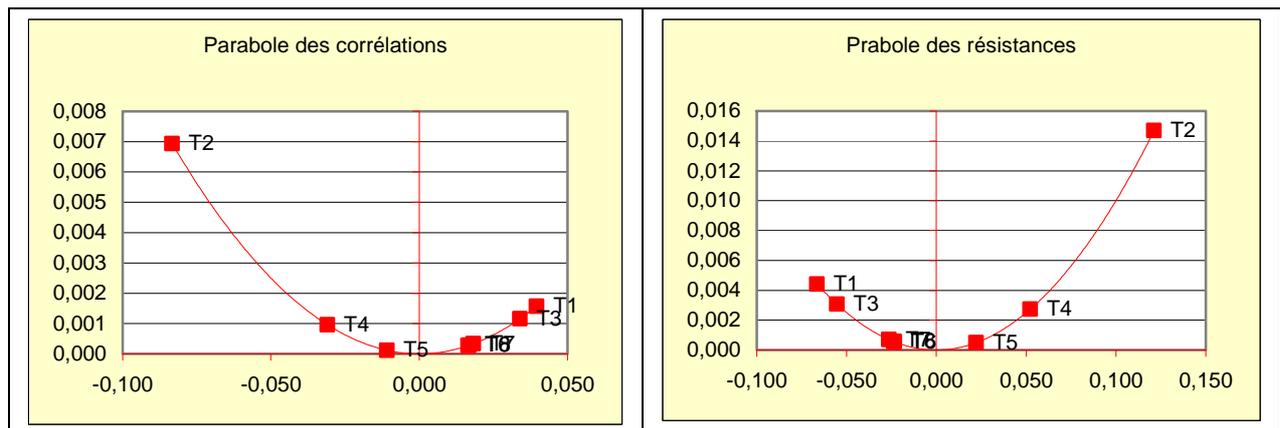
On choisira donc judicieusement les images adaptées à l'étude des phénomènes, celles qui montrent les caractéristiques fondamentales de la matrice des corrélations, *celles qui font aussi voir, revoir ou entrevoir les qualités essentielles des liaisons qui régissent le corpus.*

5.4.4 Symétrie des paraboles des résistances et des corrélations

Il en est de même des paraboles polynomiales du second degré.

Les paraboles d'équation $y = x^2$ (polynôme du second degré) sont plus expressives encore de par la puissance quadratique due à l'inertie totale du *cosinus* r'' ou du *sinus* ρ'' .

On voit des paraboles symétriques, toutes proportions gardées, (des *résistances* pour les *sinus* et des *corrélations* pour les *cosinus*) qui ne doivent leur puissance d'expression qu'à la différence de graduation des échelles, mesure de la courbe parabolique (qui projette l'algèbre dans la géométrie) :



Le tout est de bien choisir les schémas qui illustrent au mieux les caractères étudiés, mais il est inutile de les multiplier.

NB : On prendra la parabole pour illustrer les procédés de discrimination ou de lemmatisation opérés par la « Règle ». On inscrira en tête du schéma le titre de l'opération. On verra alors que la parabole polynomiale remplace avantageusement « la courbe en cloche » de Laplace-Gauss.

6. Conclusion

Il n'est guère besoin de grands discours pour savoir de quoi il s'agit lorsqu'on parle d'ACP, et de quoi il retourne lorsqu'il s'agit de choisir parmi les schémas les mieux adaptés à l'étude des phénomènes contrôlés ou résiduels.

Disons d'un mot, que l'ACP vise à produire des nuages de points dans le seul but de *monter les lieux et les distances géométriques* qui séparent ou qui rapprochent les composants de la matrice de corrélation, *une matrice carrée* par excellence.

Des schémas identiques montrant que *le degré de liaison* du nuage de points est à l'image du *degré de corrélation* des variables. Ce faisant, ils invitent à approfondir la question.

Or, la statistique se fonde sur les modèles mathématiques appropriés à la description et à la mesure des phénomènes qu'on ne peut ni voir ni mesurer *a priori*, et dont l'essence même échappe à l'ACP.

Si la statistique fournit une description des phénomènes et propose une interprétation des caractéristiques inhérentes au corpus étudié, alors elle remplit son rôle.

Aussi l'AFD prend-elle le relais de l'ACP pour identifier les éléments, les caractères et approfondir les caractéristiques, définir l'essence et la qualité des composantes, dévoiler les règles (inhérentes et intrinsèques) qui régissent tout corpus.

Aussi, en guise de synthèse, pourrions-nous citer Aristote pour qui toute « *définition est une formule qui exprime l'essentiel de l'essence du sujet* » :

Une définition est une formule qui exprime l'essentiel de l'essence d'un sujet. On peut donner, soit une formule comme l'équivalent d'un mot unique, soit une formule comme l'équivalent d'une autre formule ; de fait, il n'est pas impossible de donner des définitions de certaines choses déjà désignées par une formule. En revanche, il est bien clair que ceux qui donnent comme définition un mot unique, de quelque façon qu'ils s'y prennent, ne donnent pas une définition de ce qui les occupe, puisque précisément **une définition a toujours l'aspect d'une formule.** (...) **en matière de définitions, la discussion tourne la plupart du temps sur une question d'identité ou de différence.** (Voir Aristote, *Les Topiques*, I, 6).

L'art de la statistique que de s'appuyer sur la complémentarité de l'ACP et de l'AFD :

1. d'abord, la statistique analytique montre méthodiquement. Elle *fait voir* pour *faire savoir*, et *fait connaître* pour *faire reconnaître*. C'est le rôle de l'ACP.
2. ensuite, elle permet d'*identifier les êtres*, d'*en mesurer les rapports*, d'*en évaluer les qualités* et d'*en comprendre l'essence*. C'est le rôle de l'AFD.
3. enfin, elle répond à *toute question d'identité ou de différence* inhérente à la nature même des êtres. C'est le rôle essentiel de l'*Analyse*, qui englobe la méthode et l'objet, mettant l'une au service de la connaissance de l'autre.
4. Les conclusions ne lui appartiennent pas, elles sont d'un autre ordre et d'une nature toujours synthétique. C'est une affaire de *jugement synthétique*. Est-ce un jugement de valeur ? Est-il objectif ou subjectif ? Fondé, justifié et justifiable, vérifié et vérifiable ?

« Une théorie est d'autant plus 'scientifique' qu'elle fait entrevoir plus de *vérification* – qu'elle indique de résultats *vérifiables*. Une théorie n'est pas *vraie* ou non, elle est *vérifiable* ou non. », *dixit* Paul Valérie in *Cahiers II, La Pléiade*, page 858.

Et il ajoute plus loin : « J'appelle *crapuleux* le critique qui exploite les conclusions sans référence aux prémisses et aux opérations qui engendrent les conclusions », *ibidem*, p. 1210.

Qui dit mieux ? La méthode est disponible. Gageons que le jeu en vaut la chandelle.