Guide pratique

pour l'analyse d'épreuves ou de tests, à l'usage des étudiants et des chercheurs en Orthophonie

Illustré d'exemples avec le logiciel Hector

Quatrième version revue et augmentée,
Automne 2011
Alain Dubus
Maître de Conférences honoraire en Sciences de l'Education,
ancien professeur de statistiques à l'Institut d'Orthophonie de Lille

Sommaire

Exposé des motifs

L'auteur a enseigné l'analyse informatisée des données et les méthodes de recherche en Sciences Humaines pendant vingt-cinq années, à l'Université de Lille III. Tout au long de cette période, il a été régulièrement frappé du contraste entre le sérieux et l'ardeur développés par les étudiants et les apprentis chercheurs dans le recueil de leurs données, d'une part, et d'autre part la relative platitude des traitements mis en œuvre ensuite et des analyses qu'ils autorisaient, ce phénomène donnant envie de s'écrier « Tant d'efforts pour si peu de résultats ! ». La réponse à ce constat a consisté à tirer les enseignements de statistiques vers leurs applications pratiques, et à développer des logiciels d'analyse de données permettant une effective autonomie du chercheur débutant, par opposition à la relation frustrante avec un technicien ou ingénieur spécialiste des coûteuses « usines à gaz » que sont les grands logiciels de statistiques, mais relativement indifférent à la problématique spécifique des recherches menées.

Cette préoccupation ancienne et constante s'est colorée d'un jour nouveau pendant les quelques années, en fin de carrière, où l'auteur a été chargé d'enseigner les statistiques à des orthophonistes, d'abord en formation continue dans le cadre du DUERFO[1], puis auprès des étudiants de 3ème et de 4ème année en formation initiale. C'est précisément en préparant ces derniers au mémoire de 4ème année, spécifiquement pour celles et ceux dont le projet impliquait une approche quantitative, que l'auteur a réalisé que parmi la jungle luxuriante des instruments statistiques, les besoins réels de ces étudiants se ramenaient pour l'essentiel à un sous-ensemble relativement restreint et cohérent, que ce guide s'emploie à exposer de manière aussi pratique que possible[2].

L'objet essentiel de ce guide est de baliser quelques-uns des chemins les plus probables parmi ceux qu'auront à parcourir ses lecteurs. Même si certaines notions spécifiques sont expliquées au vol, il ne s'agit pas d'un manuel de statistiques, tant l'auteur s'est concentré sur « à quoi ça sert ? », et « comment on fait ? » plutôt que sur « Comment ça se justifie théoriquement ? ». Cependant, on trouvera en annexe un glossaire des différents termes statistiques employés, pour que l'ignorance de ces termes ne forme pas obstacle à la lecture. De plus, la plupart des exemples fournis ont été calculés avec le logiciel Hector[3] sur des données réelles provenant de mémoire d'Orthophonie récents, dont les auteurs sont ici remerciés. La connaissance d'Hector, un minimum d'autonomie dans son maniement et une certaine familiarité avec les notions de tri, de croisement et de corrélation constituent des prérequis non à la compréhension de cette notice, mais à son utilisation concrète. On trouvera en annexe une présentation rapide des fonctionnalités du logiciel ainsi qu'un cours d'initiation statistique.

---

[1] Diplôme Universitaire pour l'Enseignement et la Recherche en Orthophonie, proposé pour quelques promotions par l'Institut d'Orthophonie de Lille, et destiné aux orthophonistes professionnelles participant aux enseignements de l'Institut et aux projets de recherche de l'ARREO.

[2] Parallèlement aux tâches d'enseignement, l'expérience de l'auteur en matière de recherche quantitative en orthophonie s'est considérablement enrichie de plusieurs travaux en « vraie grandeur » à partir de 2002, comme le traitement et l'analyse des données de l'enquête « Dépistage et suivi d'enfants à risques de difficultés scolaires dès 3 ans 10 mois » (D. CRUNELLE, A. DUBUS, M-C. DUBUS, G. LICOUR, M-F. GODON), la validation et l'étalonnage d' « ECLA, outil d'évaluation des compétences langagières des enfants de 3 ans et 6 mois inclus à 6 ans et 6 mois exclu », (A. DUBUS, M-C. DUBUS, M-P LEMOINE, P. LESAGE) et, plus récemment, une contribution à la validation et à l'étalonnage de divers volets d'EVALO 2.6, à la demande de F. COQUET et J. ROUSTIT, ainsi qu'un travail de synthèse statistique sur DPL3, à la demande de F. COQUET.

[3] Sauf avis contraire, la plupart des fonctionnalités d'analyse de données utilisées ici sont disponibles dans la version de base d'Hector, à partir de la livraison 070307 du 7 Mars 2007. Certaines fonctionnalités nouvelles utilisées dans la réécriture 2008 appartiennent à Hector², versions diffusées depuis le Printemps 2008. La réécriture 2011 inclut les modifications intervenues depuis, et correspond à la refonte de la documentation d'Octobre 2011.

Références et position du problème

On aborde principalement dans ce guide des questions relatives à la validation et à l'étalonnage des tests.

Validation des tests et épreuves

On peut distinguer quatre aspects de la validation, associés à quatre vagues d'opérations distinctes, même si certaines d'entre elles peuvent être menées parallèlement. Le vocabulaire employé ici n'est pas nécessairement admis partout, mais les personnes concernées feront aisément le lien avec la nomenclature qui leur est la plus familière.

1) Validation sémantique :

Elle concerne la pertinence des items[4], c'est-à-dire, sommairement, le fait que les items aient réellement un rapport avec ce qu'ils prétendent estimer. L'objectif est de s'assurer que les items définis sont effectivement des indicateurs pertinents au regard des compétences sous-jacentes que l'épreuve est supposée estimer. Cette validation sémantique intervient en amont de toute passation, mais il arrive qu'elle fasse l'objet de révisions après les phases suivantes, dans un processus cyclique. Elle peut faire appel à diverses méthodes : consultation de professionnels, pré-tests sur le public visé… les méthodes les plus élaborées organisent la collecte et la synthèse de jugements d'experts[5].

Cet aspect de la validation n'est pas abordé dans la présente notice

2) Validation contextuelle :

Il s'agit de définir les protocoles et situation de passation auprès des publics-cibles et les conditions institutionnelles d'accès (face à face en lieu scolaire ou non, petit ou grand collectif, passation par spécialiste ou passation déléguée, consignes détaillées, contraintes temporelles …). L'objectif est de s'assurer de l'applicabilité concrète des protocoles et d'en fixer les conditions régulières de reproduction, de manière à réduire autant que faire se peut les biais de situation.

Cet aspect de la validation n'est pas non plus abordé dans la présente notice

3) Validation statistique interne :

Dite parfois validation de construct, il s'agit de l'étude des qualités métriques et statistiques des items et de leur agencement en diverses superstructures (subtests, tests, compétences, domaines, épreuves …) : difficulté, pouvoir discriminant, cohérence, fiabilité, analyses de parenté, chrono-cohérence. Les analyses proposées constituent un ensemble strict et fonctionnel largement suffisant pour fournir des appréciations robustes sur la qualité métrique des épreuves, des outils d'acceptation ou de rejet d'items et des pistes de restructuration des épreuves. Des techniques plus approfondies existent, mais réclament des connaissances statistiques spécialisées[6], ce qui n'est pas le cas des techniques proposées ici..

Cet aspect de la validation fait l'objet de la section A de cette notice

---

[4] On tentera par la suite une définition plus rigoureuse du mot item, mais on admettra pour l'instant qu'il s'agit de la plus petite entité pouvant faire l'objet d'une mesure séparée dans un épreuve.

[5] Citons entre autres auteurs Ebel, Nedelsky, Angoff, Jaeger. Ces méthodes font l'objet d'une présentation détaillée dans DE LANDSHEERE, V. (1988) *Faire réussir, faire échouer ; la compétence minimale et son évaluation*, Paris, PUF, et d'une approche plus résumée dans DUBUS, A. (2006) *La notation des élèves. Comment utiliser la docimologie pour une évaluation raisonnée*. Paris, Armand Colin.

[6] L'ouvrage de référence est ici LAVEAULT, D. & GREGOIRE, J., (1997) *Introduction aux théories des tests en Sciences Humaines*, De Boeck, Bruxelles

4) Validation statistique externe :

Dite aussi validation externe de convergence, elle s'appuie sur la comparaison des résultats de l'épreuve avec ceux obtenus, par les mêmes individus et à la même époque, à d'autres épreuves déjà validées, ou avec des observations directes, ou des résultats scolaires … voire, dans certains cas, avec les résultats de la re-passation ultérieure de la même épreuve (re-test et fidélité). Cette validation statistique externe fournit *ex post* une confirmation de la validation sémantique initiale, et une réitération de la démarche globale peut repartir de ce point.

L'analyse des qualités prédictives (prédictibilité, spécificité et sensibilité) d'une épreuve s'apparente, en termes d'outils employés, à une validation externe, même si les mesures de comparaison ont lieu à une époque ultérieure, la prédiction s'appliquant par définition à des phénomènes qui ne sont pas encore accessibles au moment où elle est émise[7]. Les instruments utilisés sont les outils ordinaires de l'étude des croisements de variables : χ^2, coefficient **r** de Bravais-Pearson, **F** de Snédécor-Fisher[8].

Cet aspect de la validation fait l'objet de la section B de cette notice

Etalonnage des épreuves et seuils critiques

L'étalonnage ne fait pas à proprement parler de la validation d'une épreuve, et à vrai dire il suppose que tous les cycles de validation nécessaires soient achevés, car l'étalonnage d'une épreuve non validée est dépourvue de sens.

L'étalonnage, qui consiste à substituer à un score brut un repère dans une échelle quantilée ou normalisée, vise à permettre la comparaison rapide et virtuelle des performances d'un individu avec celles de la population sur laquelle il a été réalisé. Les outils de l'étalonnage sont disponibles en standard dans Hector[9].

La question des seuils critiques est apparentée à la précédente : il s'agit de déterminer en dessous de quelles valeurs minimales il y a lieu d'émettre une alarme concernant un individu en considération de son score à un test.

Les questions de l'étalonnage et des seuils critiques font l'objet de la section C de cette notice. Une section D rassemble enfin diverses questions qui ne rattachent pas clairement aux précédentes.

---

[7] *La prédiction est un art difficile, notamment en ce qui concerne l'avenir*. Pierre Dac.

[8] Pour une référence « savante » et très complète, voir HOWELL, D.C. (1998) *Méthodes statistiques en Sciences Humaines*, De Boeck Université, Paris pour la traduction française. On pourra aussi se contenter des manuels d'Hector en ligne, surtout *Hector_manuel_traitements 2011.pdf* et *Hector_annexe_statistique.pdf*. On peut aussi se reporter en priorité à l'Annexe 3 de ce guide, qui rassemble les éléments les plus importants des cours de statistiques dispensés naguère à l'Institut d'Orthophonie de Lille.

[9] De plus, l'ouvrage cité DUBUS, A. (2006) consacre la majeure partie de son onzième chapitre à l'analyse de cette pratique et de ses effets en docimologie.

Section A : la validation statistique interne

Vocabulaire et préalables

item

C'est la plus petite entité susceptible de recevoir une mesure, d'être affectée d'une valeur ; l'item correspond le plus souvent à une question ou à une réalisation, mais parfois une question peut engendrer plusieurs items.

Pour qu'il soit aisé d'effectuer des travaux statistiques sur l'épreuve, il vaut mieux qu'aux items correspondent des mesures numériques, et mieux encore des mesures binaires (0 ou 1, faux ou vrai). Parfois, on peut être amené à dichotomiser (c'est-à-dire ramener au binaire) une mesure plus étendue au moyen d'un seuil de coupure.

superstructures

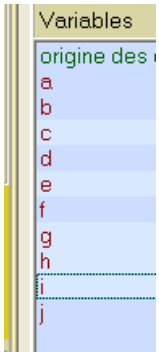
C'est tout ce qui organise les items en ensembles plus ou moins étendus.

L'organisation la plus fréquente est en trois niveaux : item / sub-test / test, ou encore item / épreuve / test. L'idée sous-jacente, souvent implicite en dépit de ses implications méthodologiques, est que le niveau intermédiaire (sub-test ou épreuve) rassemble des items cohérents entre eux, qui tendent à prélever des indices variés de la même chose, de quoi qu'il s'agisse : compétence, habileté, connaissance… le niveau le plus élevé rassemblant au contraire plusieurs objets de niveau intermédiaire nettement différenciés entre eux. Le constructeur de tests et d'épreuves a intérêt à décider clairement si c'est bien ce modèle qu'il adopte.

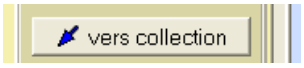
Des organisations de superstructures plus complexes sont envisageables : ainsi les tests d'évaluation CE2 utilisent quatre niveaux, item / compétence / domaine / test, et en Français un niveau composante s'intercale entre item et compétence.

Dans Hector les items sont représentés par des *variables*, et les superstructures sont représentées par des *collections*, ensembles de variables créés par la décision de l'opérateur en vue d'effectuer des opérations – précisément – collectives.

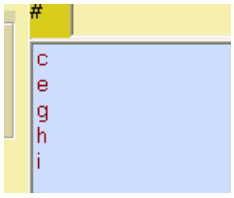
La manœuvre de création d'une collection dans Hector est aisée, dès qu'on dispose des variables utiles. Dans la page nommée VARIABLES, dans le panneau des variables en haut à droite,



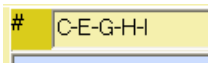
on sélectionne les variables que l'on veut grouper, avec des clic, des majuscule-clics et des contrôle-clics si besoin est, comme dans toute liste Windows. Les variables sélectionnées apparaissent sur un fond plus clair.



on clique le bon bouton

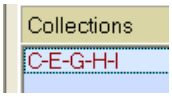


les variables viennent se ranger où il faut



on tape un nouveau nom de collection

on valide en bas à droite



et la nouvelle collection est désormais disponible.

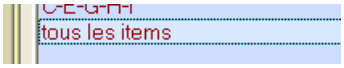
Disponible pour quoi faire ? Pas mal de choses. En premier lieu il est facile d'écrire une formule (dans la page Formules du logiciel) qui créera une nouvelle variable contenant la somme des éléments d'une collection numérique. Si on agit ainsi en créant une collection pour chaque compétence du test CE2, on peut faire calculer un score pour chaque compétence, puis on peut passer ensuite au niveau supérieur en créant des collections de ces scores de compétence, à raison par exemple d'une par domaine, et on peut en faire calculer la somme, puis faire une collection des scores totaux de domaines et enfin faire calculer un score global… Problème : ce n'est pas parce qu'il facile de faire des additions qu'il est légitime et judicieux de le faire, comme on l'apprend aux petits enfants (on n'additionne pas des poireaux et des carottes !) mais comme on s'empresse de l'oublier quand on devient une grande personne (en calculant des moyennes scolaires générales par exemple).

Difficulté et discrimination

Une première vérification intéressante à opérer sur un ensemble d'items regroupés dans une collection concerne la difficulté de ces items, qui est le complément du taux de réussite.

On va dans la page Traitements d'Hector, on s'assure que l'onglet Collectifs est actif.

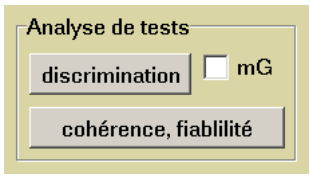
On double clique dans la liste des collections sur le nom de la collection qu'on veut étudier (ici une bien nommée « tous les items »).



ça l'installe dans la boîte à étudier les collections :

on aurait pu, aussi, simplement la sélectionner, puis cliquer le bouton ajouter ; ça revient au même.

Plus haut, le bouton discrimination est disponible et prêt à l'emploi :



et ceci, parce que tous les items de la collection sont binaires, c'est-à-dire qu'ils ne peuvent avoir que les valeurs 0 ou 1 (des items logiques, Vrai/faux, feraient aussi bien l'affaire).

Sinon, par exemple avec des items contenant des mesures d'une étendue plus vaste, le bouton serait grisé et inutilisable, car seuls les items binaires permettent ces calculs. C'est une des raisons pour lesquelles on recommande d'avoir des items à score binaire quand c'est possible, ou d'en fabriquer au moyen de seuils dans les autres cas.

En cliquant le bouton discrimination, on obtient le tableau suivant :

```
Analyse de la collection logique ou binaire {tous les items}
taux de réussite, difficulté
indice de discrimination et qualité
```

REUSSITE		faibles	médians	forts	DISCRIM	qualité	item
0,26	-	0,11	0,25	0,51	0,40	Ok	a
0,82	+	0,63	0,89	0,93	0,30	-	b
0,62	=	0,24	0,71	0,95	0,70	Ok**	c
0,52	=	0,23	0,53	0,91	0,68	Ok**	d
0,29	-	0,10	0,22	0,73	0,63	Ok**	e
0,18	-	0,09	0,14	0,40	0,32	Ok	f
0,22	-	0,07	0,16	0,56	0,48	Ok*	g
0,63	=	0,25	0,72	0,96	0,71	Ok**	h
0,52	=	0,15	0,59	0,86	0,71	Ok**	i
0,70	+	0,45	0,75	0,92	0,47	Ok*	j

La première colonne est celle du taux de réussite. Il est exprimé ici comme une fréquence entre 0 et 1, qui équivaut à l'intervalle 0% à 100%.

Le système de signes qui commente le taux de réussite est arbitrairement inspiré de l'expérience :

-- de 0 à 15%

- de 16 à 35%

= de 36 à 64%

+ de 65 à 84%

++ de 85 à 100%

Quelle est l'importances des taux de réussite ? Eh bien un item très difficile ou très facile n'est pas très intéressant du point de vue de l'information qu'il peut apporter, car l'analyse statistique vise à faire apparaître des différences entre les cas et les situations ; de ce point de vue, les taux de réussite les plus intéressants en termes de théorie de l'information sont autour de 50% : ce sont ceux qui facilitent les tâches de comparaison et de classement.

Un test dont tous les items seraient trop difficiles ou trop faciles n'est tout simplement pas adapté à la population étudiée. On notera au passage (on y reviendra) que la difficulté

d'un item n'a de sens que vis-à-vis d'une population donnée : ce qui est difficile pour vous ne l'est pas nécessairement pour moi, et *vice versa*.

Les trois colonnes suivantes sont encore des taux de réussite, mais calculés pour des sous-groupes distincts d'individus : les *faibles*, les *médians* et les *forts*.

Comment sont constitués ces groupes ? Sur la base du score obtenu en totalisant les scores obtenus aux items. Les faibles ont les 27% de plus faibles au regard de ce score total, les forts sont les 27% de plus forts, et les médians les 46% qui restent au milieu[10].

Ainsi, pour l'item *a*, le taux de réussite des faibles est de 11%, celui des médians de 25%, celui des forts de 51%.

L'intérêt principal de la manœuvre réside dans la cinquième colonne, qui contient l'indice de discrimination de chaque item : c'est la différence entre le taux de réussite des forts et celui des faibles.

Un indice de discrimination élevé dénote un item utile dans la mesure où l'on souhaite construire un test qui sépare clairement les individus selon des niveaux de performance contrastés. A l'inverse, un item qui est plus ou moins réussi, mais à peu près autant par les faibles et les forts, ne traite probablement pas des mêmes compétences que les autres items de l'épreuve.

La colonne qualité du tableau expose des appréciations sur la qualité discriminante des items, de manière à permettre le repérage rapide des anomalies :

Ok++ correspond à une discrimination d'au moins 0,50, qu'on pourrait qualifier d'excellente

Ok+ va de .40 à .49, c'est très satisfaisant

Ok va de .30 à .39, c'est correct

- va de .20 à .29, c'est faible : l'item n'est à conserver que si on ne peut absolument pas s'en passer (par exemple parce qu'on s'est laissé enfermer dans une situation où on n'en n'a pas de rechange).

-- va de .10 à .19 : il est préférable d'éviter d'utiliser un tel item

?? est pour les items dont l'indice de discrimination tombe en dessous de .10 : ceux-là n'apportent rien à l'épreuve

!!! est pour les items aberrants, par exemple ceux que les faibles réussissent mieux que les forts.

Bien qu'il s'agisse d'un indice formel et insensible à la signification, il va de soi que des indices de mauvaises qualité remettent en cause la construction de l'épreuve et/ou son adaptation au public visé.

On prendra également garde au fait que les indices de discrimination sont plus faibles quand l'épreuve contient un nombre important d'items (20 ou 30), parce qu'avec un nombre plus petit d'items, la présence de l'item lui-même dans le total qui permet de définir les trois classes d'individus tend à biaiser les résultats en faveur de l'item ; dans le cas de petits nombres d'items, comme ici, il faut donc être plus exigeant[11].

Comme pour le taux de réussite, les indices de discrimination sont à référer à la population étudiée. Le tableau ci-dessous étudie les items de la compétence « Mots fréquents » du domaine « Ecriture – Orthographe » du test ce2 Français en 2005 dans une circonscription du Nord de la France (plus de 500 enfants):

[10] Ces valeurs arbitraires, proposées par Kelley et reprises par Findley, correspondent environ à 0.61 écarts-types de part et d'autres de la moyenne d'une distribution normale.

[11] C'est pour anticiper cette exigence que nous avons ajouté une classe OK++ au-delà de 0,40, alors qu'Ebel ne la prévoyait pas.

REUSSITE		faibles	médians	forts	DISCRIM	qualité	item
0.95	++	0.83	0.97	1.00	0.17	--	F28
0.61	=	0.27	0.63	0.89	0.62	Ok**	F29
0.61	=	0.10	0.64	0.99	0.89	Ok**	F30
0.89	++	0.55	0.96	1.00	0.45	Ok*	F88
0.72	+	0.16	0.82	0.99	0.83	Ok**	F89
0.63	=	0.24	0.63	0.98	0.74	Ok**	F90
0.44	=	0.03	0.39	0.95	0.92	Ok**	F91
0.27	-	0.01	0.18	0.70	0.69	Ok**	F92
0.44	=	0.04	0.39	0.91	0.88	Ok**	F93

La plupart des items discriminent de manière très satisfaisante, voire exceptionnelle, à l'exception de F28, qui est aussi le plus facile avec 95% de succès. Les autres items affichent des taux de réussite non extrêmes, sauf le F88 très facile lui aussi, mais bien discriminant quand même.

Le tableau change quand on considère la même compétence, mais en restreignant[12] le champ de l'étude à la grosse centaine d'enfants « en retard » :

```
Analyse de la collection logique ou binaire {EcrOrtho mots fréquents}
taux de réussite, difficulté
indice de discrimination et qualité
sous le filtre en retard
```

REUSSITE		faibles	médians	forts	DISCRIM	qualité	item
0.88	++	0.69	0.97	0.97	0.28	!!!	F28
0.55	=	0.25	0.61	0.81	0.56	Ok**	F29
0.44	=	0.06	0.50	0.75	0.69	Ok**	F30
0.85	+	0.53	1.00	1.00	0.47	Ok*	F88
0.57	=	0.08	0.71	0.92	0.83	Ok**	F89
0.45	=	0.14	0.39	0.83	0.69	Ok**	F90
0.25	-	0.03	0.00	0.75	0.72	!!!	F91
0.13	--	0.03	0.05	0.31	0.28	-	F92
0.29	-	0.03	0.29	0.56	0.53	Ok**	F93

Les taux de réussite ont descendu de plusieurs points, mais deux items, F28 et F91, sont devenus aberrants, tandis que F92, devenu très (trop ?) difficile, n'est pratiquement plus utilisable. On prendra donc soin, quand on a des raisons de suspecter des disparités importantes au sein de la population étudiée, d'analyser séparément les diverses composantes : l'outil ne convient pas également à tous ; on retrouvera des préoccupations similaires en matière d'étalonnage, par exemple avec des enfants d'âge différent.

Récupérer les tableaux affichés

Comme partout dans Hector, cliquer le bouton -> Document à définir ouvre une boîte de dialogue permettant à l'usager de donner un nom de fichier de type .RTF (utilisable par Word) dans lequel seront recopiés les tableaux et graphiques affichés, chaque fois qu'on cliquera ce bouton, qui affichera désormais le nom du fichier. Pour utiliser ce fichier sous Word, il faut d'abord le fermer sous Hector, soit en droite-cliquant le bouton, soit en quittant Hector.

[12] Grâce à la technique du *filtre*, qui permet de définir temporairement un sous-ensemble de la population, sur lequel les traitements sont effectués.

Une technique plus rapide, implémentée après 2008, permet de copier-coller tableaux et graphiques directement depuis Hector vers un logiciel de traitement de texte du commerce (Microsoft Word) ou libre (Libre Office).

L'option « modèle de Guttman »

Si la case mG est cochée, cela indique qu'on souhaite confronter la collection au modèle de Guttman.

Celui-ci s'applique normalement à une collection d'items binaires mesurant à des niveaux divers une même compétence, ou une même dimension de compétence. Le principe est que, si un sujet a réussi un item d'un certain niveau de difficulté, on s'attend à ce qu'il ait réussi aussi à tous les items de difficulté inférieure.

Ainsi, avec quatre items de difficulté croissante a, b, c, d, le profil de réussite 1 1 0 0 est conforme au modèle, puisqu'il est celui des individus qui ont réussi a et b, mais ni c ni d.

En revanche, un profil 1 0 1 0 n'est pas conforme, puisque les individus concernés n'ont pas réussi b, alors qu'ayant réussi c ils auraient « dû » réussir aussi b.

Les seuls profils acceptables sont les suivants :

0 0 0 0, 1 0 0 0, 1 1 0 0, 1 1 1 0 et 1 1 1 1

Ils ont en commun qu'aucun 1 ne doit apparaître à la *droite* d'un 0, et, réciproquement, aucun 0 à *la gauche* d'un 1.

Si l'on arrange les profils conformes au modèle dans un tableau approprié, on obtient le tableau suivant :

a	b	c	d
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

La forme en escalier justifie le nom de modèle pyramidal parfois donné à cette forme. Il a des chances de se produire dans un système ou chaque capacité plus rare englobe les précédentes.

Le coefficient de reproductibilité[13] de Guttman est le quotient du nombre de cases convenables par le nombre total de cases (nombre de sujets x nombre de variables) dans le grand tableau à une ligne par sujet et une colonne par variable. Une case non convenable est une case qui contient un 0 (échec) à la gauche d'un 1 (réussite à un item réputé plus difficile). L'autre mesure est le pourcentage de sujets qui présentent des profils rigoureusement compatibles avec le modèle de Guttman. La fréquence croissante des erreurs permet d'identifier dans quelle mesure chaque variable contribue au nombre total d'erreurs.

```
Analyse de la collection logique ou binaire pas de problèmes % phrases
taux de réussite, difficulté
indice de discrimination et qualité

Reproductibilité (modèle de Guttman) : 0,83
% de sujets rigoureusement conformes = 59,01 (théorique 31,25)
seuils de confiance à .01, .001, .0000 : 36,05 37,62 39,50
fréquence croissante des erreurs :
0,00 pas de pb % phrases
```

[13] Nous conservons l'appellation traditionnelle, dont on pourrait toutefois discuter la pertinence.

```
0,18 pas de pb % phrases exp
0,32 pas de pb % phrases m.f
0,51 pas de pb % phrases +2t
```

L'exemple ci-dessus porte sur la mesure de la complexité syntaxique dans le PER 2000 (échantillon PRS/ARREO campagne initiale) : taux de phrases, de phrases à expansion, à module fonctionnel, avec les deux extensions. Chaque variable « pas de pb… » indique que, selon l'étalonnage de Ferrand-Nespoulous pour son âge, le score de l'enfant est considéré comme normal.

La reproductibilité de .83 est relativement importante, mais pas assez pour qu'on puisse se fier entièrement au modèle : on exige usuellement pour cela un coefficient de .90. La ventilation des erreurs par variable peut permettre à un stade de la mise au point du test, si le modèle pyramidal est souhaité, quel(s) item(s) il faudrait exclure pour améliorer le coefficient de Guttman.

Alors que le coefficient de Guttman fait plutôt porter la « responsabilité » des erreurs sur les items, le taux de sujets conformes mesure à quel point la distribution observée s'éloigne du taux théorique de sujets conformes[14] sous l'hypothèse de réponses indépendantes. Ce n'est pas une mesure très exigeante en soi que l'éloignement du modèle aléatoire. Aussi utilise-t-on des seuils de décision très fins : .01, .001, .0000 (quasi certitude). La question à laquelle il est répondu n'est pas « Est-ce que cette collection présente une structure pyramidale ? », mais « A quel point le nombre de sujets conformes au modèle de la structure pyramidale s'écarte-t-il de ce que le hasard aurait pu provoquer ? »

Dans l'exemple, le taux de conformité supérieur à 59% permet d'écarter l'hypothèse nulle (alétoire) : il y a bien une tendance à la structure pyramidale, mais elle n'est pas parfaite (coefficient de Guttman à .83). Un taux de conformité aussi significatif pourrait conduire à rechercher pour quelles parties de la population étudiée le modèle serait mieux satisfait… mais ceci nous éloigne du propos principal.

Cohérence et fiabilité

La seconde vérification des qualités métriques des épreuves et tests concerne la cohérence d'une épreuve, ou sub-test. Le modèle sous-jacent est que l'épreuve est constituée d'items parallèles en contenu et en difficulté, qui constituent autant d'indicateurs imparfaits mais convergents d'une compétence sous-jacente, polluée par le « bruit[15] » de la situation de test.

Cette notion de cohérence est extrêmement importante d'un point de vue pratique, car elle légitime le fait de procéder à des additions de scores d'items pour produire un score d'épreuve. Pour le dire plus familièrement, une cohérence élevée garantit qu'on additionne bien des poireaux avec des poireaux et non avec des carottes.

On y accède en cliquant le bouton cohérence, fiabilité, juste en dessous du bouton discrimination , la collection à étudier étant dans la boîte à étudier les collections. Si l'on vient d'étudier la discrimination, elle y est encore.

```
Analyse de la collection numérique ou logique {tous les items}
_____________________________________________________________________
     moyenne    écart-type       r(i,T-i)        item
        0,26          0,44         0,118          a
```

[14] Une collection de n variables binaires engendre n+1 profils conformes, parmi les 2^n profils possibles. Le taux théorique dans l'exemple, avec quatre variables, est de 5/16, soit 31,25%.

[15] *Bruit* par opposition au *signal*. Ici le bruit représente les éléments perturbateurs tels que stress, inattention, réponses au hasard, fatigue).

```
0,82        0,38        0,160         b
0,62        0,48        0,386***      c
0,52        0,50        0,312***      d
0,29        0,45        0,290***      e
0,18        0,38        0,127         f
0,22        0,41        0,234**       g
0,63        0,48        0,385***      h
0,52        0,50        0,301***      i
0,70        0,46        0,201*        j

Corrélations item/test (cohérence) : min, moy, max : 0,118, 0,251, 0,386
Alpha de Cronbach (fiabilité) = 0,572
```

La première colonne reprend le score moyen à l'item, qui équivaut au taux de réussite quand la variable est binaire (cette fonctionnalité est également accessible aux items non-binaires). La seconde colonne fournit l'écart-type de ce score (indice de dispersion).

La troisième colonne contient la véritable mesure de cohérence : la corrélation item-test. Elle est ainsi appelée dans le sens ou le test serait la seule superstructure à l'item, mais évidemment il s'agit d'une corrélation item-subtest ou item-épreuve.

Plus précisément, il s'agit, pour chaque item, de mesurer la corrélation entre l'item lui-même et la somme des items de l'épreuve, l'item lui-même exclu. C'est ainsi qu'il faut comprendre le titre un peu sybillin de la colonne : r(i,T-i) ; *r* est mis pour corrélation entre *i*, l'item et *T-i*, la somme des items, sans l'item considéré.

Quand cette corrélation est élevée, cela signifie que l'item est bien à sa place dans cette épreuve, qu'il contribue efficacement à constituer la mesure globale que sera la somme des scores aux items, autrement dit le score à l'épreuve : une forte cohérence légitime le fait même de calculer un tel score par addition de scores partiels.

Certaines corrélations portent une, deux ou trois astérisques *. Avec *, c'est une corrélation significative au seuil de .10, avec **, au seuil de .05, avec ***, au seuil de .01. Sans signe, c'est non significatif.

Qu'en est-il de l'exemple ? Eh bien, ce n'est pas fameux. Quand on sait que l'effectif concerné frôle les 500 sujets, il y a lieu d'être exigeant, et des corrélations à .300, même très significatives grâce aux effectifs abondants[16], ne sont pas pour autant très fortes.

Force des corrélations

En effet, le carré de la corrélation r^2 entre deux variables peut être interprété comme la part d'information commune à ces deux variables. Ainsi une corrélation à .300 signifie que la part d'information commune aux deux variables est de .09, soit moins de 10%.

Piéron[17] qualifie ainsi de très faibles des corrélations de l'ordre de .200, de faibles des corrélations autour de .300, de moyennes celles de .400 à .500, et d'assez élevées celles de .800, laissant entendre qu'un seuil qualitatif interviendrait vers .700, ce qui ne laisse quand même qu'un r^2 proche de .50, ou 50%.

[16] De manière générale, la significativité des tests est d'autant plus fine, à relation équivalente, que l'effectif est important, raison pour laquelle, en présence d'effectifs réduits (n<50), on peut accepter des seuils P<.05, voire P<.10, ce qui serait une faute avec des effectifs plus importants (n>100), où il faut être plus exigeant.

[17] PIERON, H. (1969), *Examens et docimologie*, 2e éd., Paris PUF. Cet ouvrage fondateur de la docimologie en France reprend en fait des travaux menés avant la seconde guerre mondiale, et interrompus par celle-ci.

Dans le cas étudié, ce n'est donc pas très brillant. Tout au plus serait-il envisageable de constituer une assez médiocre[18] épreuve réduite aux items c, d, e, h et i, qui ont des corrélations faibles au sens de Piéron.

Analyse de la collection numérique ou logique {cdehi}

moyenne	écart-type	r(i,T-i)	item
0,62	0,48	0,378***	c
0,52	0,50	0,288***	d
0,29	0,45	0,291***	e
0,63	0,48	0,357***	h
0,52	0,50	0,292***	i

Corrélations item/test (cohérence) : min, moy, max : 0,288, 0,321, 0,378
Alpha de Cronbach (fiabilité) = 0,560

En éliminant les items les plus « mauvais », on approche par endroits les corrélations moyennes au sens de Piéron.

La ligne en dessous du tableau reprend dans la liste de corrélations item-test le minimum, la moyenne et le maximum.

La dernière ligne présente la statistique alpha de Cronbach , qui est une mesure de fiabilité. Ce n'est pas une corrélation, mais une estimation de la probabilité que les items mesurent la même chose, que les erreurs se compensent pour que la somme délivre la mesure d'une valeur sous-jacente. Ici, des valeurs comme 0,572 ou 0,560 sont très insuffisantes : il n'y a pratiquement pas plus d'une chance sur deux qu'un tel modèle soit réaliste. Un alpha de 0,750 paraît un minimum, et 0,900 est très bon[19].

On peut améliorer la statistique de fiabilité alpha de Cronbach en augmentant le nombre d'items, à condition que ceux-ci soient au moins aussi cohérents[20] que ceux qui existent déjà. La formule suivante[21] :

$$k = [\ a_1\ (1 - a_0)\ ]\ /\ [a_0\ (1 - a_1)\ ]$$

où a_0 désigne l'alpha de Cronbach actuel, et a_1 l'alpha de Cronbach souhaité, fournit k, coefficient par lequel il faut multiplier le nombre actuels d'items pour espérer atteindre la fiabilité souhaitée (toujours sous la condition d'items cohérents).

En partant de l'épreuve {cdehi} et avec comme objectif une fiabilité à 0,750[22],

---

[18] Les pitoyables qualités métrologiques de ce test ne doivent rien au hasard : il s'agit en effet d'un test « pour rire », un toy-problem fabriqué exprès pour mettre en évidence certains phénomènes docimologiques dans A. DUBUS (2006) *op. citat.*, p. 48 « Gammes sur une épreuve inventée »

[19] Certains auteurs mettent le seuil d'acceptabilité à 0,700, d'autres à 0,800.

[20] L'ajout d'items non cohérents n'améliore pas grand-chose : si on lit, dans l'autre sens, le passage de cdehi à la liste complète des items, le gain de fiabilité n'est que de 0.560 à 0.572.

[21] Dûe à Spearman-Brown, cf. LAVEAULT, D. & GREGOIRE, J., (1997), *op. citat*, p.154. Le même ouvrage développe ensuite des considérations sur le calcul de l'erreur-type sur le score global, utilisant le taux de fiabilité. Cette erreur-type permet ensuite de calculer l'intervalle de confiance d'un score global donné, à différents seuils de probabilité. Ces éléments très utile pour la comparaison de populations entre elles ou pour l'estimation de la vraisemblance de l'appartenance d'un individu à une population nous ont paru cependant situés au-delà de la qualification statistique exigée pour utiliser avec bonheur le logiciel Hector dans l'analyse des tests.

[22] Ce seuil de 0,750 s'apprécie ordinairement sur un test ou sub-test constitué d'une dizaine d'items. Ajouter des items cohérents augmente cette fiabilité, en enlever la diminue. Pour permettre la comparaison entre tests dotés d'un nombre différents d'items, Hector² affiche aussi l' « alpha comparable pour 10 items », qui mesure au moyen de la formule ci-dessus quel serait

k = [0,75 × (1- 0,56)] / [0,56 × (1 – 0,75)], soit k = 2,36

Il faudrait donc passer à une douzaine d'items de même qualité pour atteindre une fiabilité minimale. Pour atteindre une fiabilité de 0,900 (90%), il faudrait multiplier le nombre d'items par 7 !

Parentés et structures

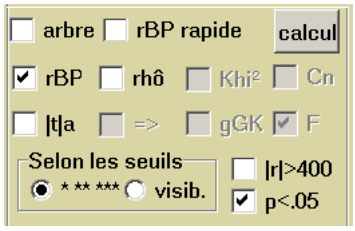
Il s'agit maintenant d'étudier les relations que les items, ou que les entités de rang supérieur, entretiennent entre elles, et si ces relations sont de nature à mettre en cause l'agencement des superstructures sur les items (subtests ou épreuves, mais aussi compétences, domaines …).

Matrice des corrélations

Une collection étant installée dans la boîte à étudier les collections, il faut maintenant l'installer une seconde fois, soit en re-double-cliquant son nom, soit avec le bouton ajouter alors que le nom de la collection et sélectionné dans la liste.

Ce dispositif qui peut sembler étrange est dû au fait que la Matrice de statistiques, dont il va être fait usage ci-après, permet de croiser entre elles deux collections différentes ou une seule collection avec elle-même (ce qui nous intéresse ici).

Dès que deux collections sont sélectionnées, la Matrice de statistiques peut fonctionner : le bouton calcul devient fonctionnel, et les différentes options de la matrice deviennent disponibles dans la mesure où le type des variables le permet.



S'agissant d'une collection numérique à croiser avec elle-même, la statistique proposée par Hector est le coefficient de corrélation r de Bravais-Pearson, ou rBP. On pourrait utiliser aussi le coefficient rhô de Spearman, qui est la même chose mais sur les rangs ; dans le cas de variables binaires les résultats sont identiques.

La statistique |t|a, test de Student sur échantillons appareillés, est hors sujet ici : elle sert essentiellement à évaluer des effets avant/après.

Collection M1.3 × elle-même

Matrice des coefficients de corrélation r (Bravais-Pearson)

	itemMath49	itemMath50	itemMath51	itemMath52	itemMath53
itemMath49		0.716 ***	0.202 *	0.166	0.187 *
itemMath50	0.716 ***		0.200 *	0.161	0.210 **
itemMath51	0.202 *	0.200 *		0.659 ***	0.383 ***
itemMath52	0.166	0.161	0.659 ***		0.356 ***
itemMath53	0.187 *	0.210 **	0.383 ***	0.356 ***	

l'alpha si on amenait le nombre d'items à 10, *à qualité de cohérence sous-jacente constante*. Ce n'est donc pas une modalité alternative du calcul de l'indice de fiabilité, mais un simple outil de comparaison.

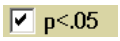
Les données utilisées ici sont extraites des résultats de l'évaluation CE2 en 2005 sur 938 enfants dans deux circonscriptions du Nord-Pas-de-Calais.

Les items de Mathématiques 49 à 53 constituent ensemble la troisième compétence « Comparer les nombres entiers naturels au moyen de graduations » du premier champ « Nombre naturels », d'où le code M1.3

Chaque item se retrouve sur une ligne et sur une colonne. La valeur à l'intersection ligne x colonne dans le tableau est la corrélation entre l'item de la ligne et l'item de la colonne. Le tableau ne comporte pas de valeurs sur sa diagonale, car la corrélation d'une variable avec elle-même est par définition égale à 1 (l'identité)[23]. Le tableau est symétrique autour de sa diagonale parce la corrélation est elle-même une statistique symétrique : r(x,y)=r(y,x), ce qui n'est pas forcément le cas pour d'autres statistiques.

Les corrélations portent *** si elles sont significatives à .01, ** à .05, * à .10, et rien du tout si elles ne sont pas significatives. Rappelons qu'à côté de la significativité (probabilité qu'avait le simple hasard de fournir de telles valeurs), il faut considérer la force de la corrélation (faible vers .300, forte vers .700).

Usuellement, pour faciliter la lecture des matrices de corrélation, on n'affiche pas les valeurs qui ne sont pas ou trop peu significatives. Cela s'obtient en cochant la case d'option[24] :



et en relançant le calcul[25].

Collection M1.3 × elle-même

Matrice des coefficients de corrélation r (Bravais-Pearson)

	itemMath49	itemMath50	itemMath51	itemMath52	itemMath53
itemMath49		0.716 ***			
itemMath50	0.716 ***				0.210 **
itemMath51				0.659 ***	0.383 ***
itemMath52			0.659 ***		0.356 ***
itemMath53		0.210 **	0.383 ***	0.356 ***	

Que lit-on ? Des corrélations assez fortes à fortes entre les items 49 et 50 d'une part, 51 et 52 d'autre part. L'item 53 entretient pour sa part des corrélations très faibles à plutôt faibles avec les items 50 à 52, mais aucune corrélation significative avec l'item 49. Comment interpréter ces éléments ? En indiquant qu'on a deux paires d'items très solidaires (49-50 et 51-52), et un cinquième item plus distant (53).

[23] Hector ne l'affiche pas, pour améliorer la lisibilité.

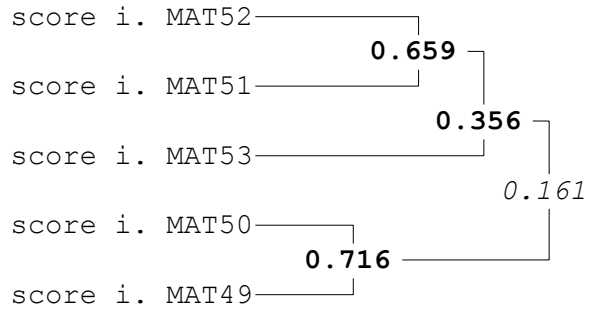
[24] Le tableau présenté ici est obtenu avec l'option « selon les seuils * ** *** ». Il est également possible de choisir l'option « visib. », qui proposera des seuils plus fins, jusqu'à .0000, et n'affichera que les corrélations significatives à ce seuil, ce qui est très utile quand on recherche les corrélations les plus saillantes parmi un grand nombre de variables.

[25] D'autres options sont disponibles, mais sans utilité immédiate dans notre propos. L'option rBP rapide s'applique quand on est sûr que toutes les variables concernées ont une valeur définie pour chaque sujet, elle accélère le traitement pour les très grands ensembles de données. L'option |r|>400 inhibe l'affichage des corrélations inférieures à .400 en valeur absolue (seuil des corrélations consistantes au sens de Piéron).

Arborescence des parentés

La même information peut se retrouver dans l'arborescence des parentés, qu'on demande avec la case à option appropriée arbre, avant de relancer le calcul :

On obtient le schéma suivant :



Cela ressemble vaguement à un arbre couché, dont le tronc serait à droite, et dont les feuilles, à gauche, sont les items.

La démarche de construction de cette arborescence consiste à rechercher les deux items les mieux corrélés, et à les assembler : ici les items 50 et 49. Assemblés, il constituent une entité dont on détermine la corrélation avec les autres entités, dont les items encore isolés. Cette corrélation avec la nouvelle entité est, par construction, la plus petite corrélation constatée avec un des éléments de cette entité.

Cela paraît un peu abstrait, mais, si on regarde l'assemblage suivant, celui des items 51 et 52 : il porte sa corrélation 0.659. Ensuite, l'item 53, qui est corrélé à 0.383 avec l'item 51, mais à 0.356 avec l'item 52, prend pour corrélation avec ce groupe de deux items la plus petite corrélation des deux, soit ici 0.356.

Cela a pour conséquence que la corrélation portée au sommet d'un groupe d'items est en quelque sorte la corrélation minimale garantie entre deux quelconques des membres du groupe.

Les corrélations écrites en **gras** sont significatives à .01, celles écrites en caractères ordinaires sont significatives à .05, celles écrites en *italique* sont significatives à .10 ou pas du tout.

Ceci confirme, en l'affinant, le diagnostic précédent : deux paires d'items très liés(49-50, 51-52), un cinquième lié plus lâchement (53), pas d'unité d'ensemble. Cette compétence M1.3 manque d'unité, et c'est sans doute fâcheux pour l'interprétation de ses résultats.

De manière générale, on notera que l'arborescence des parentés repose sur un critère très exigeant, nettement plus que la cohérence-fiabilité, et qu'elle ne doit donc pas être utilisée pour remettre en cause cette dernière, mais plutôt pour étudier les apparentements dans un grand nombre d'items, par exemple dans le but de les organiser selon des groupements non prévus initialement[26].

[26] La version professionnelle-recherche d'Hector propose des outils plus sophistiqués pour étudier la connectivité des ensembles d'items, mais leur usage exige de s'approprier une démarche assez complexe, qui n'intéressera que les chercheurs confirmés.

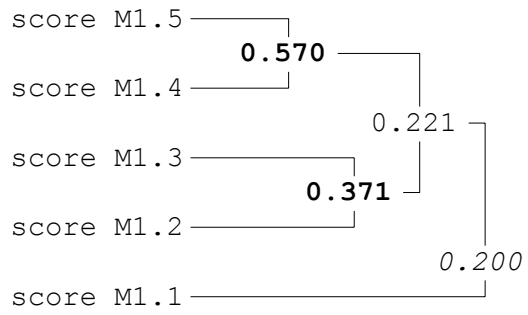
A un niveau supérieur de la structure

On a ici analysé une compétence, qui selon notre nomenclature courante constitue un subtest ou une épreuve. Si l'échafaudage structurel comporte plus de deux niveaux, on peut répéter l'opération à un niveau supérieur, par exemple avec le champ (ou domaine) M1 nombres naturels :

Collection Scores M1 Nombres naturels × elle-même

Matrice des coefficients de corrélation r (Bravais-Pearson)

	ScoreM11	ScoreM12	ScoreM13	ScoreM14	ScoreM15
ScoreM11		0.225 **			0.274 ***
ScoreM12	0.225 **		0.371 ***	0.221 **	0.461 ***
ScoreM13		0.371 ***		0.251 **	0.432 ***
ScoreM14		0.221 **	0.251 **		0.570 ***
ScoreM15	0.274 ***	0.461 ***	0.432 ***	0.570 ***	



On retrouve à nouveau deux paires bien liées : M1.5-M1.4, et à un degré moindre M1.2-M1.3. Toutefois l'absence de fortes intercorrélations entre tous les éléments est ici moins problématique qu'entre items, et peut-être même qu'au contraire il est judicieux que des compétences diversifiées soient mesurées par des scores faiblement corrélés entre eux. Tout dépend du genre d'outil qu'on veut construire, et des possibilités d'interprétation qu'on en espère.

Dans le cas des tests d'évaluation CE2, une analyse systématique laisse à penser que la nomenclature des activités et des objets sur lesquels elles portent a pris ou conservé le pas sur une logique des compétences, non directement observables et en quelque sorte transversales, car une analyse de tous les items de mathématiques pris ensemble, qu'on épargnera ici au lecteur, montre que les assemblages issus des corrélations ne correspondent que rarement au découpage en compétences, domaines et champs. Il en est de même en Français[27]. Ceci dit, ces tests n'ont jamais prétendu à l'excellence métrique, et, destinés à faciliter la communication entre enseignants et leur hiérarchie, il était logique qu'y prévale le vocabulaire usuel de la profession et les découpages du réel qu'il sous-tend.

[27] Sophie Morlaix, dans *Identifier et évaluer les compétences dans le système éducatif : quels apports pour la recherche en éducation*. Rapport d'habilitation à diriger les recherches, 2007, montre d'ailleurs que certaines compétences transversales se retrouvent dans des items de mathématiques, mais aussi de français.

Une étude de cas de validation interne sur un test orthophonique

L'étude de cas qui suit est basée sur le travail réel de deux étudiantes[28] en quatrième année d'orthophonie, dont la tâche consistait à valider un test de repérage de difficultés en CE1, et à définir un seuil critique sur ce test.

Pour ce faire, elles avaient élaboré un test de huit épreuves couvrant quatre domaines :

- Conscience phonologique
  - conscience phono 1 : comptage de syllabes dans des mots entendus
  - conscience phono 2 : présence / absence d'un son dans des mots entendus
- Orthographe
  - orthographe 1 : dictée de logatomes
  - orthographe 2 : dictée de mots existants
- Lecture
  - lecture 1 : lecture de logatomes
  - lecture 2 : lecture de mots existants

Elles entreprennent la vérification systématique des qualités métriques de leur test, passé par 97 enfants de CE1.

```
Analyse de la collection logique ou binaire conscience phono 1
taux de réussite, difficulté
indice de discrimination et qualité
```

REUSSITE		faibles	médians	forts	DISCRIM	qualité	item
0.96	++	0.85	1.00	1.00	0.15	--	SEAU
0.94	++	0.77	1.00	1.00	0.23	-	ASPIRATEUR
0.98	++	0.92	1.00	1.00	0.08	??	HERISSON
0.99	++	0.96	1.00	1.00	0.04	??	BOUGIE
0.84	+	0.38	1.00	1.00	0.62	Ok**	ANANAS
0.94	++	0.77	1.00	1.00	0.23	-	ORDINATEUR

```
Analyse de la collection numérique ou logique conscience phono 1
```

moyenne	écart-type	r(i,T-i)	item
0.96	0.20	-0.022	SEAU
0.94	0.24	0.428***	ASPIRATEUR
0.98	0.14	0.154	HERISSON
0.99	0.10	0.269**	BOUGIE
0.84	0.37	0.164	ANANAS
0.94	0.24	0.016	ORDINATEUR

```
Corrélations item/test (cohérence) : min, moy, max : -0.022, 0.168, 0.428
Alpha de Cronbach (fiabilité) = 0.312
```

La conformité à un modèle pyramidal n'étant pas recherchée, on n'emploie pas l'option mG .

Le diagnostic sur la première épreuve est assez sévère (et le choc est rude). Aucun item n'a de qualité discriminante, à part « Ananas », un seul offre une corrélation item-test significative et décente, le coefficient de fiabilité de Cronbach est très faible.

[28] Remerciements à ANNE DEPREY ET CLAIRE RENARD, (2007) *Mémoire présenté pour l'obtention du certificat de Capacité d'Orthophoniste*, Institut d'Orthophonie Gabriel Decroix, Université de Lille II.

La cause de cet échec apparent réside très vraisemblablement dans une erreur de calibrage de la difficulté de l'épreuve au regard du public visé : les items sont tous très faciles, et un item trop facile ne peut pratiquement jamais être discriminant : le plus difficile, qui est aussi le seul items discriminant, « Ananas », est réussi par 84% des enfants !

Il est décidé d'abandonner complètement cette épreuve, qui ne peut rapporter aucune information utilisable.

L'épreuve suivante, Conscience phonologique 2, apporte des résultats plus rassurants :

Analyse de la collection logique ou binaire conscience phono 2
taux de réussite, difficulté
indice de discrimination et qualité

REUSSITE		faibles	médians	forts	DISCRIM	qualité	item
0.68	+	0.23	0.74	1.00	0.77	Ok**	ch
0.76	+	0.33	0.87	1.00	0.67	Ok**	z
0.52	=	0.10	0.35	1.00	0.90	Ok**	k
0.84	+	0.60	0.87	1.00	0.40	Ok	i
0.78	+	0.37	0.90	1.00	0.63	Ok**	a
0.66	+	0.20	0.71	1.00	0.80	Ok**	b

Analyse de la collection numérique ou logique conscience phono 2

moyenne	écart-type	r(i,T-i)	item
0.69	0.46	0.499***	ch
0.75	0.43	0.631***	z
0.52	0.50	0.599***	k
0.84	0.37	0.382***	i
0.77	0.42	0.570***	a
0.67	0.47	0.568***	b

Corrélations item/test (cohérence) : min, moy, max : 0.382, 0.542, 0.631
Alpha de Cronbach (fiabilité) = 0.788

Les items, tous faciles mais pas trop, sauf un moyen (k), offrent tous au moins une bonne qualité discriminante et le plus souvent une excellente. Les corrélations item-test sont satisfaisantes (sauf « i », un peu médiocre), et le coefficient de fiabilité est très correct, surtout pour un si petit nombre d'items. Une recherche d'excellence pourrait conduire à essayer d'améliorer l'item « i », et même à essayer d'ajouter de nouveaux items (d'autant que « Conscience phonologique 1 » a coulé corps et biens), mais l'organisation logistique de l'opération et les délais ne semblent pas le permettre.

L'épreuve Conscience phonologique 2 est donc conservée sans modifications.

L'épreuve suivante, « Orthographe 1 », pose un problème : les scores peuvent être de 0, 1 ou 2, et les items ne sont pas des variables binaires.

Cela est gênant à plusieurs égards.

Tout d'abord, question de commodité, des variables non-binaires ne permettent pas d'évaluer l'indice de discrimination.

Par ailleurs, le motif de cette notation en trois valeurs est que les logatomes à orthographier comportent deux syllabes, et que l'erreur peut intervenir une ou deux fois. Ce souci de précision est intéressant mais inopérant : en effet le fait de coder le nombre de réussites empêche de distinguer les erreurs sur la première syllabe des erreurs sur la seconde, ce qui eût pu présenter un intérêt analytique détaillé : c'est assez typiquement

le genre de choses auxquelles il faut avoir pensé avant, car on ne sait pas décomposer après coup un score additionné dès la saisie.

L'examen des distributions des scores à ces items montre que la médiane passe le plus souvent entre 2 (aucune erreur) et 1 (une seule erreur), les 0 (deux erreurs étant assez rares).

Il est donc décidé de dériver[29] de ces items de nouvelles variables valant 0 s'il y a au moins une erreur et 1 s'il n'y en a aucune. Grâce à cette modification, on a de nouveau accès à la discrimination :

```
Analyse de la collection logique ou binaire ² orthographe 1
taux de réussite, difficulté
indice de discrimination et qualité

_____________________________________________________________________
REUSSITE   faibles  médians   forts   DISCRIM  qualité   item
0.84  +      0.46     0.96     1.00     0.54     Ok**     ² trali
0.40  =      0.08     0.34     0.84     0.76     Ok**     ² gorbu
0.49  =      0.12     0.49     0.88     0.76     Ok**     ² séjou
0.72  +      0.35     0.81     0.96     0.61     Ok**     ² ronca
0.69  +      0.23     0.81     0.96     0.73     Ok**     ² fémou
0.76  +      0.23     0.91     1.00     0.77     Ok**     ² nouvi
0.61  =      0.23     0.64     0.96     0.73     Ok**     ² muchon
0.69  +      0.35     0.74     0.96     0.61     Ok**     ² dapon
0.61  =      0.19     0.64     1.00     0.81     Ok**     ² siclu
0.49  =      0.15     0.47     0.88     0.73     Ok**     ² jétol
Analyse de la collection numérique ou logique ² orthographe 1
_____________________________________________________________________
     moyenne   écart-type      r(i,T-i)       item
        0.84         0.37       0.552***      ² trali
        0.40         0.49       0.420***      ² gorbu
        0.49         0.50       0.422***      ² séjou
        0.72         0.45       0.460***      ² ronca
        0.69         0.46       0.546***      ² fémou
        0.76         0.43       0.623***      ² nouvi
        0.61         0.49       0.473***      ² muchon
        0.69         0.46       0.435***      ² dapon
        0.61         0.49       0.599***      ² siclu
        0.49         0.50       0.465***      ² jétol
Corrélations item/test (cohérence) : min, moy, max : 0.420, 0.500, 0.623
Alpha de Cronbach (fiabilité) = 0.815
```

Tous les items modifiés sont faciles ou moyens, leurs indices de discrimination sont tous excellents. De même les corrélations item-test sont très significatives et de force moyenne (minimum 0.420) ou mieux. Le coefficient de fiabilité alpha de Cronbach est tout à fait satisfaisant.

L'épreuve aux scores modifiés « ² Orthographe 1 » est donc conservée sans autres modifications.

[29] Au moyen du langage des formules.

L'épreuve « Orthographe 2 » est composée d'items binaires, et ne réclame donc pas de modification préalable :

```
Analyse de la collection logique ou binaire orthographe 2
taux de réussite, difficulté
indice de discrimination et qualité

_____________________________________________________________________
REUSSITE   faibles  médians   forts   DISCRIM  qualité   item
0.50  =      0.04     0.53     0.95     0.91     Ok**     ballon
0.70  +      0.24     0.80     1.00     0.76     Ok**     chat
0.56  =      0.16     0.61     0.91     0.75     Ok**     cinq
0.29  -      0.00     0.22     0.77     0.77     Ok**     cuisine
0.51  =      0.16     0.51     0.91     0.75     Ok**     eau
0.55  =      0.28     0.51     0.95     0.67     Ok**     sept
0.56  =      0.12     0.61     0.95     0.83     Ok**     orange
0.64  =      0.20     0.73     0.95     0.75     Ok**     maison
0.56  =      0.16     0.63     0.86     0.70     Ok**     poison
0.14  --     0.00     0.18     0.23     0.23      -       coq
Analyse de la collection numérique ou logique orthographe 2
_____________________________________________________________________
     moyenne   écart-type      r(i,T-i)       item
        0.49         0.50       0.494***      ballon
        0.70         0.46       0.536***      chat
        0.55         0.50       0.417***      cinq
        0.28         0.45       0.473***      cuisine
        0.51         0.50       0.340***      eau
        0.54         0.50       0.309***      sept
        0.55         0.50       0.557***      orange
        0.63         0.48       0.521***      maison
        0.56         0.50       0.444***      poison
        0.14         0.35       0.221**       coq
Corrélations item/test (cohérence) : min, moy, max : 0.221, 0.431, 0.557
Alpha de Cronbach (fiabilité) = 0.767
```

Un item pose un important problème de discrimination : il s'agit de « coq », item très difficile[30], et assez logiquement insuffisamment discriminant. De plus, le même item n'offre qu'une corrélation item-test faible. L'élimination de cet item s'impose. Dans une perspective d'amélioration de l'épreuve, les items 'eau' et 'sept' pourraient sans doute être remplacés.

Cependant le coefficient de fiabilité et déjà satisfaisant, et l'ablation de l'item défaillant le fait monter à 0.770, en dépit de la diminution du nombre d'items.

On utilisera donc par la suite une épreuve « Orthographe 2 (9 items) » amputée de l'item « coq ».

L'épreuve suivante, « Lecture 1 » est composée d'items non binaires. Il est donc procédé à la même dérivation que pour « Orthographe 1 ».

[30] L'auteur se souvient douloureusement d'avoir commis sur ce mot sa seule faute d'orthographe à la dictée de l'examen d'entrée en sixième, vers 1955. En se relisant, un coup de panique hypercorrectrice lui a fait ajouter un 'u', au motif que le 'q' ne saurait achever un mot français (et « cinq », alors ?). Fatalitas !

Analyse de la collection logique ou binaire ² lecture 1
taux de réussite, difficulté
indice de discrimination et qualité

REUSSITE	faibles	médians	forts	DISCRIM	qualité	item
0,76 +	0,44	0,79	0,97	0,53	Ok**	² prafi
0,57 =	0,36	0,39	0,91	0,55	Ok**	² vouga
0,75 +	0,48	0,76	0,94	0,46	Ok*	² loutu
0,71 +	0,44	0,66	0,97	0,53	Ok**	² chédon
0,74 +	0,32	0,82	0,97	0,65	Ok**	² caqué
0,90 ++	0,72	0,95	0,97	0,25	-	² mori
0,79 +	0,36	0,89	1,00	0,64	Ok**	² sonju
0,62 =	0,12	0,63	0,97	0,85	Ok**	² borchi
0,67 +	0,24	0,76	0,88	0,64	Ok**	² flécou
0,57 =	0,16	0,53	0,91	0,75	Ok**	² punon

Analyse de la collection numérique ou logique ² lecture 1

moyenne	écart-type	r(i,T-i)	item
0,77	0,42	0,371***	² prafi
0,57	0,49	0,187*	² vouga
0,76	0,43	0,361***	² loutu
0,71	0,45	0,328***	² chédon
0,73	0,44	0,478***	² caqué
0,90	0,30	0,289***	² mori
0,80	0,40	0,544***	² sonju
0,62	0,48	0,560***	² borchi
0,66	0,47	0,487***	² flécou
0,57	0,49	0,457***	² punon

Corrélations item/test (cohérence) : min, moy, max : 0,187, 0,406, 0,560
Alpha de Cronbach (fiabilité) = 0,743

Deux problèmes de nature différente apparaissent : l'item « mori », trop facile, est clairement non-discriminant (et de plus faiblement corrélé item-test), et l'item « vouga », bien discriminant, est en revanche très mal corrélé item-test.

Les deux items sont donc supprimés. Le coefficient de fiabilité remonte alors à 0.756, ce qui n'est pas encore extraordinaire. Il apparaît nettement qu'il eût été plus satisfaisant de remplacer « mori » et « vouga « que de les supprimer.

Dans une perspective d'amélioration du test, « prafi », « loutu » et « chédon » auraient peut-être pu être enlevés aussi, pour cause de faible corrélation item-test, si du moins une solution de remplacement avait existé.

On utilisera donc par la suite l'épreuve modifiée « ² Lecture 1 (8 items) » amputée des items « mori » et « vouga ».

La dernière épreuve, « Lecture 2 » est constituée d'items binaires et ne nécessite donc pas de modification préalable.

Analyse de la collection logique ou binaire lecture 2
taux de réussite, difficulté
indice de discrimination et qualité

REUSSITE		faibles	médians	forts	DISCRIM	qualité	item
0.86	++	0.44	1.00	1.00	0.56	Ok**	vert
0.87	++	0.60	0.93	1.00	0.40	Ok	loup
0.64	=	0.32	0.61	1.00	0.68	Ok**	nez
0.77	+	0.20	0.93	1.00	0.80	Ok**	histoire
0.90	++	0.60	1.00	1.00	0.40	Ok	je suis
0.70	+	0.40	0.70	1.00	0.60	Ok**	trop
0.93	++	0.80	0.96	1.00	0.20	--	dans
0.89	++	0.68	0.93	1.00	0.32	Ok	oiseau
0.51	=	0.00	0.50	1.00	1.00	Ok**	clown
0.61	=	0.16	0.63	1.00	0.84	Ok**	monsieur

Analyse de la collection numérique ou logique lecture 2

moyenne	écart-type	r(i,T-i)	item
0.86	0.35	0.663***	vert
0.87	0.34	0.491***	loup
0.65	0.48	0.337***	nez
0.77	0.42	0.671***	histoire
0.90	0.30	0.521***	je suis
0.71	0.46	0.375***	trop
0.93	0.26	0.417***	dans
0.89	0.31	0.483***	oiseau
0.51	0.50	0.561***	clown
0.61	0.49	0.481***	monsieur

Corrélations item/test (cohérence) : min, moy, max : 0.337, 0.500, 0.671
Alpha de Cronbach (fiabilité) = 0.809

En raison du faible pouvoir discriminant de l'item « dans », il est décidé de l'éliminer.

On utilisera donc par la suite l'épreuve modifiée « lecture 2 (9 items) »

Les cinq épreuves subsistantes possèdent désormais des qualités suffisantes pour qu'on puisse calculer un score somme sur chacune d'elles. Ces cinq scores sommes sont ensuite assemblés en une collection pour étudier leurs relations au niveau structurel supérieur :

Analyse de la collection numérique ou logique scores épreuves

moyenne	écart-type	r(i,T-i)	item
4,23	1,86	0,670***	score conscience phono 2
6,31	2,85	0,806***	² score orthographe 1
4,88	2,60	0,736***	score orthographe 2 (9 items)
5,62	2,19	0,801***	²score lecture 1 (8 items)
6,74	2,30	0,772***	score lecture 2 (9 items)

Corrélations item/test (cohérence) : min, moy, max : 0,670, 0,757, 0,806
Alpha de Cronbach (fiabilité) = 0,898

Le coefficient de fiabilité est très élevé, de même que les corrélations item-test. Cela justifiera le calcul d'un score global, puisque les éléments qui le composent manifestent une grande cohérence.

En revanche, on peut s'interroger sur la spécificité des différents domaines, qu'on pourrait soupçonner de tous mesurer la même chose.

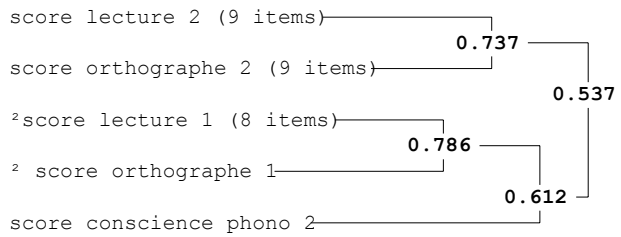
Collection scores épreuves × elle-même

Matrice des coefficients de corrélation r (Bravais-Pearson)

	ScoConPho2	²ScorOrth1	ScOr29It	²sLe18It	ScLe29It
ScoConPho2		0.612 ***	0.537 ***	0.648 ***	0.538 ***
²ScorOrth1	0.612 ***		0.611 ***	0.786 ***	0.677 ***
ScOr29It	0.537 ***	0.611 ***		0.627 ***	0.737 ***
²sLe18It	0.648 ***	0.786 ***	0.627 ***		0.649 ***
ScLe29It	0.538 ***	0.677 ***	0.737 ***	0.649 ***	

Les corrélations entre épreuves sont toutes très significatives et moyennes à importantes

L'arborescence des parentés fournit des clefs de lecture complémentaires :



Lecture 2 et orthographe 2 sont très liés, ils ont commun d'opérer sur de vrais mots.

Lecture 1 et orthographe 1 sont très liés aussi, ils ont en commun d'opérer sur des logatomes.

Conscience phonologique 2 est plus proche des épreuves portant sur des logatomes que des épreuves portant sur de vrais mots.

La corrélation minimale, 0.537, moyenne au sens de Piéron, autorise l'hypothèse d'une autonomie partielle des épreuves.

De fait, il est usuel que des épreuves portant sur des aptitudes scolaires ou des capacités langagières soient très corrélées entre elles, ce qui peut faire douter de la spécificité de chaque épreuve. C'est que dans ce genre de situation une dimension sous-jacente, qu'on peut associer selon les cas au degré général d'excellence scolaire ou au stade de développement langagier, domine et sature les épreuves particulières.

Une analyse en composantes principales permet d'illustrer ce phénomène. Il s'agit d'une analyse de type factoriel qui s'applique à des variables numériques et s'emploie, sur la base de leurs corrélations, à extraire des facteurs explicatifs successifs, orthogonaux entre eux et d'inertie (information) décroissante.

Cette technique permet d'expliquer par un nombre réduit de facteurs les relations entre un ensemble plus grand de variables.

La fonctionnalité d'analyse en composantes principales[31], de même que celle du plan de projection, ne font pas partie de la version de base gratuite, mais de la version professionnelle standard.

[31] On verra que cette technique n'apporte pas réellement d'autres informations que ce qui précède, même si elle les illustre mieux. En résumé, elle est souvent appréciée pour son côté *techno-fun*.

```
Analyse en composantes principales (préliminaire)
97 sujets sont concernés
Taux d'inertie des axes
     propre   cumulé
 1 : 71,89%   71,89%
 2 : 11,05%   82,95%
 3 :  7,67%   90,61%
 4 :  5,26%   95,87%
 5 :  4,13%  100,00%
```

Les trois premiers axes concentrent plus de 90% de l'inertie, les deux suivants sont anecdotiques

```
Corrélations des variables
axe   1     2     3     4     5
_____________________________________
    +.78  +.48  +.39  -.06  -.01   score conscience phono 2
    +.89  +.10  -.31  -.06  -.32   ² score orthographe 1
    +.83  -.39  +.22  +.31  -.08   score orthographe 2 (9 items)
    +.88  +.18  -.29  +.17  +.29   ²score lecture 1 (8 items)
    +.85  -.35  +.04  -.36  +.12   score lecture 2 (9 items)
```

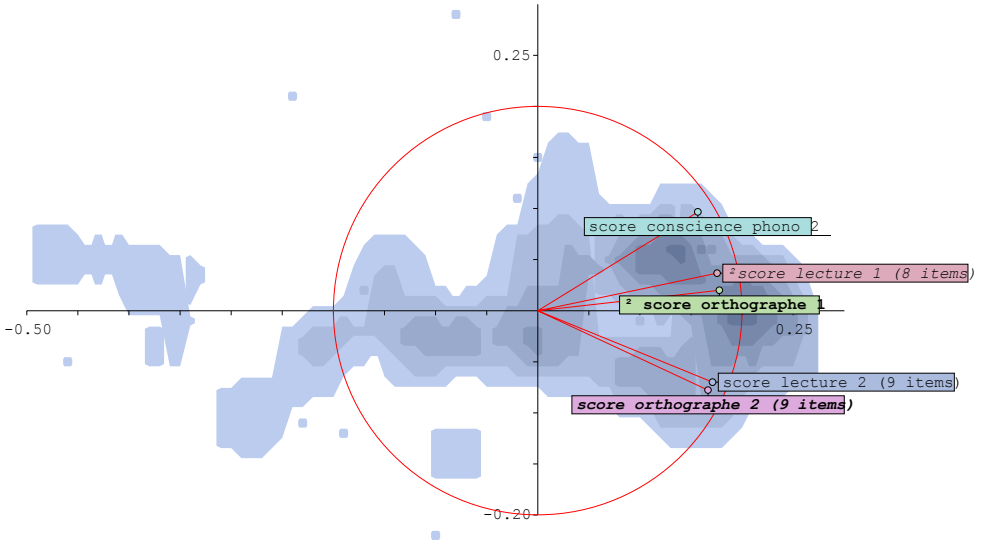
Les cinq variables d'origine sont uniformément corrélées très fortement avec le premier axe, qui s'annonce donc comme un facteur général très puissant. Les contrastes n'apparaissent qu'avec les axes suivants.

```
Carré des corrélations des variables (*100)
axe   1   2   3   4   5  tous
_____________________________
     61  23  15   0   0  100   score conscience phono 2
     79   1  10   0  10  100   ² score orthographe 1
     69  15   5  10   1  100   score orthographe 2 (9 items)
     77   3   8   3   8  100   ²score lecture 1 (8 items)
     73  12   0  13   1  100   score lecture 2 (9 items)
```

Le carré des corrélations des variables avec les axes successifs montre comment l'inertie de chaque variable se partage entre les axes. On observera (en calculant un peu) que l'inertie propre de chaque axe (premier tableau) n'est autre que la moyenne des carrés des corrélations de cet axe avec chaque variable : les 71,89 % d'inertie du premier facteur sont (aux arrondis près) la moyenne des carrés de corrélations sur l'axe 1 : 61 , 79 , 69, 77, 73, tous à diviser par 100 bien sûr.

L'étape suivante va consister à projeter, sur un plan dont les deux axes sont précisément les deux premiers axes de l'ACP, les variables d'origine et le nuage représentant les sujets.

```
Plan de projection :
en x, Acp sur scores épreuves axe1
en y, Acp sur scores épreuve axe2
```



L'axe horizontal est le premier axe, le facteur général[32]. Le nuage bleuté représente les individus : il devient plus sombre là où la densité de sujets est élevée. On voit que le nuage s'étire le long du premier axe : ce n'est pas un hasard, c'est précisément l'effet mathématique recherché.

Les variables de score des épreuves constitutives sont représentés comme des vecteurs dont l'origine est au 0,0 du plan, et dont l'extrémité a pour coordonnées les corrélations des variables avec les deux axes : la corrélation avec le premier axe est aussi le cosinus de l'angle que fait le vecteur de la variable avec l'axe horizontal, la corrélation avec le second axe est aussi le sinus du même angle

Le cercle rouge est appelé « cercle des corrélations » une variable dont l'extrémité serait posée précisément sur ce cercle aurait la somme des carrés de ses corrélations = 1 (au motif que $\cos^2+\sin^2 = 1$ pour tout angle). Les cinq variables de score d'épreuves n'atteignent pas ici une telle perfection dans la projection, mais en sont bien près.

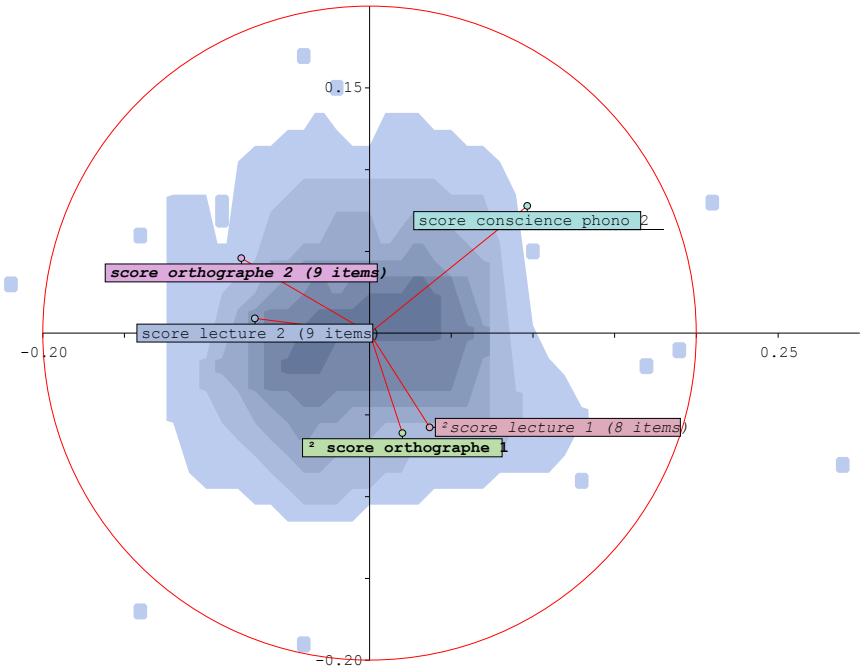
Les vecteurs des cinq variables sont tous assez allongés le long du premier axe, ce qui manifeste l'importance de ce gros facteur commun, mais manifestent par ailleurs des divergences d'orientation, avec les logatomes et la conscience phonologique en haut, opposés aux épreuves sur de vrais mots vers le bas.

Une dernière projection va permettre d'affiner cette analyse :on va projeter les mêmes éléments sur un plan constitué des axes 2 et 3. Cela signifie que l'on fait abstraction de ce qui est commun aux scores (l'axe 1), pour se concentrer sur ce qui le différencie (les axes suivants).

[32] Il est corrélé avec la somme des items aux épreuves à 0.999, c'est-à-dire pratiquement l'identité. On obtient également un facteur général de ce type quand on additionne les scores obtenus à différents modules d'un diplôme universitaire : il dénote alors la qualité académique globale des sujets.

On peut se représenter l'image suivante comme représentant une coupe au point 0,0, par un plan perpendiculaire à l'axe 1.

```
Plan de projection :
en x, Acp sur scores épreuve axe2
en y, Acp sur scores épreuve axe3
```



Cette fois, les choses sont nettes : conscience phonologique, logatomes et vrais mots partent dans trois directions distinctes du plan. La relative sympathie de la conscience phonologique pour les logatomes s'exprime sur l'axe 2 (maintenant horizontal), mais non sur l'axe 3, qui a permis de « déplier » les trois composantes du total. A ce stade du propos, il importe de se souvenir que les axes 2 et 3 ne totalisent que 18,59% de l'inertie totale, et qu'il serait donc oiseux de les commenter plus avant.

Retour sur la validation interne

Choisir ses outils

Les trois types d'analyses proposées, difficulté-discrimination, cohérence-fiabilité et parenté-structures, ont des rôles distincts mais complémentaires, et l'importance qu'on leur accorde doit être pondérée par les objectifs de la mise au point du test, selon qu'on le veut à spectre large ou au contraire concentré (en contenu), adapté à un vaste public ou au contraire spécialisé (en exigences), à structure simple ou au contraire complexe-encyclopédique.

Ainsi, la cohérence-fiabilité et de fortes corrélations item-test sont très souhaitables au niveau le plus bas de la structure, mais, si on les retrouve en comparaison de scores à plusieurs sub-tests ou épreuves, il peuvent se retourner en indices de non-spécificité de ces épreuves les unes vis à vis des autres.

Anticiper dès la conception de l'épreuve

Dans tous les cas, il est hautement recommandé de prévoir pour la phase de mise au point un nombre d'items par épreuve nettement supérieur (d'un facteur 1,5 au moins) au nombre d'items qu'on espère pouvoir conserver, de manière à pouvoir éliminer sans états d'âme les items dysfonctionnels. Se retrouver en fin de validation avec un nombre d'items insuffisant pour obtenir une fiabilité satisfaisante, ou être obligé de conserver des items

dont on a prouvé la médiocrité peut constituer un affreux gâchis des efforts déployés, situation qu'un peu de générosité dans l'élaboration initiale aurait pu éviter.

La question des âges

Les données utilisées jusqu'ici dans cette section, y compris dans l'étude de cas, concernaient des populations d'âge homogène.

Or on peut évidemment se trouver dans des situations où les sujets relèvent de tranches d'âge diverses et parfois découpées très finement, avec l'hypothèse sous-jacente que les compétences étudiées sont de nature à évoluer avec l'âge, et donc les scores à augmenter.

Ceci introduit deux types de considérations, sans même songer aux questions d'étalonnage qui seront reprises en section C :

- Il faut d'abord vérifier si cette hypothèse, dite de chrono-cohérence est juste
- Il faut éventuellement mener les analyses de discrimination séparément par groupes d'âges

Dans l'exemple suivant[33], l'hypothèse de chrono-cohérence n'est pas juste, en tous cas pas avec le degré de détail utilisé dans les âges. On croise la variables classe d'âge avec l'un des scores. S'agissant du croisement d'une variable catégorielle avec une variable numérique, ce qu'on obtient est une analyse de variance[34].

Analyse de la variance de (score mémoire kinesthésique 18 item) selon les positions de (classe d'âge)

Classe	Effectif	Moyenne	Ecart-type
3a3-3a9	11	6.82	5.44
3a9-4a3	11	10.55	7.01
4a3-4a9	10	12.70	5.10
4a9-5a3	10	13.50	5.06
5a3-5a9	10	17.40	1.80
5a9-6a3	11	16.27	3.67
ENSEMBLE	63	12.79	6.12

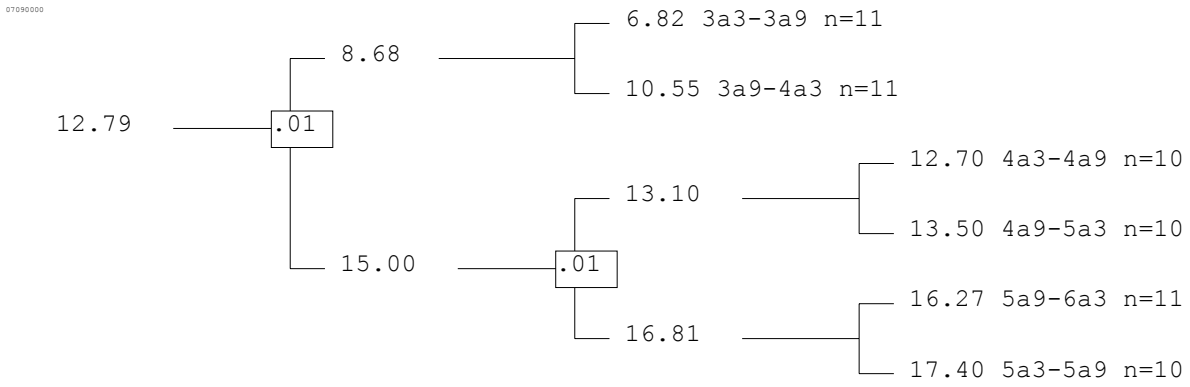
Pour chacun des groupes de 10 à 11 sujets tranches d'âge, on obtient la moyenne du groupe à ce score et son écart-type. Le sens de la démarche est de déterminer si les moyennes des différents groupes sont significativement différentes ou si l'on peut considérer qu'il ne s'agit que de fluctuations aléatoires autour de valeurs similaires.

F(5,57) = 5.82 s. à .001

Le test F de Snédécor-Fisher donne une valeur très significative (.001, c'est-à-dire une chance sur mille que ce soit dû au hasard), qui permet d'affirmer que *globalement* la classe d'âge induit certaines différences de moyenne. Peut-on dire pour autant qu'aux classes d'âge corresponde une progression des scores ?

[33] Données issues du mémoire de Ludivine BURET et Jessica NICLI (2008), *Validation d'un parcours Attention / mémoire de la batterie Evalo 2.6*. Mémoire présenté pour l'obtention du certificat de Capacité d'Orthophoniste, sous la direction de Françoise COQUET et Jacques ROUSTIT. Institut d'Orthophonie Gabriel Decroix, Université de Lille II.

[34] ANOVA dans les ouvrages qui prisent la terminologie anglo-saxonne.



L'arborescence des contrastes permet d'en douter dans le détail.

En effet, seules deux bifurcations, dotées d'un « .01 », correspondent à des contrastes significatifs[35]. Elles organisent donc les classes d'âge en trois sous-ensembles distincts :

- 3a3-3a9 et 3a9-4a3
- 4a3-4a9 et 4a9-5a3
- 5a3-5a9 et 5a9-6a3

Les différences à l'intérieur de ces sous-ensembles ne sont pas significatives selon le |t| de Student. De plus, dans le sous-ensemble des plus âgés, l'ordre des classes est inversé.

Dans cet exemple, l'hypothèse de chrono-cohérence n'est acceptable qu'à condition d'employer des classes d'âge d'un an et non de six mois.

En ce qui concerne la *discrimination* et la *difficulté*, dans le cas où la population est composée de strates d'âge contrastées, il n'est pas nécessairement indiqué d'éliminer systématiquement les items trop difficiles ou pas assez, ou encore peu discriminants, sur la base d'une analyse globale tous âges confondus, puisque certains items peuvent être difficiles pour les plus jeunes, mais standard pour de plus âgés. Il convient donc de procéder à ces analyses sous le filtre des différentes classes d'âge présentes, et de conserver ou de rejeter les items selon le critère de leur utilité discriminante dans *l'une ou l'autre* classe d'âge. Pour la cohérence-fiabilité, cette précaution n'est pas nécessaire.

[35] Dans les versions les plus récentes d'Hector, la significativité va jusqu'à .0000.

Section B : validation externe, qualités prédictives et fidélité

La validation interne ne faisait appel qu'aux données de l'épreuve elle-même, administrée en un unique temps t.

La validation externe au sens large peut se décomposer en différents cas de figure, sur les critères du temps et des instruments de mesure employés :

- Si l'on relève, en même temps que le test à valider ou dans un temps très proche, d'autres mesures (tests, observations, résultats scolaires…) supposées refléter en tout ou partie les mêmes capacités que le test lui-même, on s'attend à ce que la similitude des compétences visées soit statistiquement confirmée par des corrélations. Il s'agit alors de *validation externe* au sens strict.
- Si l'on relève à un temps t_1 ultérieur au temps t_0 de passation initiale du test en question, d'autres mesures supposées refléter des capacités apparentées, mais à un stade plus évolué, le premier test est regardé comme devant anticiper des capacités à venir, qu'il n'est pas encore possible de mesurer directement, mais dont les épreuves du test sont espérés comme signes avant-coureurs de l'évolution ultérieure, on est dans *l'étude des qualités prédictives* du test.
- Si l'on refait passer aux mêmes sujets, à un temps t_1 ultérieur au temps $t_{0,}$ le même test, il s'agit d'une mesure de la *fidélité* du test, c'est à dire de sa constance comme instrument de mesure. Cela n'a de signification que si le phénomène observé n'est pas en soi susceptible d'évolution, soit spontanée (maturation), soit provoquée (apprentissage) ; sinon, il ne s'agit plus d'études du test lui-même, mais de l'étude de l'évolution du phénomène, comme dans les études *avant/après*.

Validation externe au sens strict

Sous les conditions exposées ci-dessus (même population même temps, mesures distinctes), on va recouper les résultats du test avec d'autres résultats. Le but de l'opération est de s'assurer que le test mesure bien ce que l'on souhaite, et d'en apporter la preuve en le comparant avec une épreuve dont on sait ce qu'elle mesure. On perçoit dès cette définition les limites de l'exercice : si l'autre test mesure la même chose et qu'on lui fait confiance, pourquoi en construire un nouveau ?

Dans les faits, on ne recherche donc pas l'identité absolue des résultats, ce qui serait absurde, mais une certaine consistance :

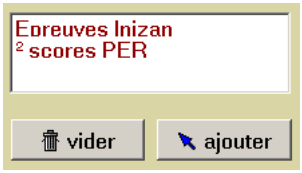
- Soit en ne comparant que les parties du test à valider et celles de la mesure témoin qui traitent effectivement de compétences similaires.
- Soit en considérant que l'on construit une test plus court et plus léger sur des compétences déjà mesurables par un protocole plus lourd et coûteux, le bénéfice matériel de l'opération pouvant justifier une certaine approximation dans l'ajustement.
- Soit enfin, comme la tendance se développe, le test plus courts est destiné à être passé systématiquement à de grandes populations comme un signal de première alerte, les formes plus lourdes n'étant passées que dans les cas les plus critiques au vu du premier test.

L'outil statistique de base de la validation externe est la corrélation, pour autant que les valeurs construites par le test à valider et le test de référence soient des variables numériques. Si les types de variables sont différents, on pourra en cas de besoin utiliser le

rhô de Spearman sur croisement d'ordinales, voire l'analyse de variance s'il s'agit de recouper une variable numérique avec une variable catégorielle[36].

On peut croiser deux par deux toutes les variables des deux tests, mais il est plus efficace d'utiliser la matrice statistique :

On installe (en sélectionnant dans la liste des collections, puis bouton ajouter) dans la boîte à collections la collection des scores du test à valider et la collection des scores du test de référence :



L'exemple du PER

On utilise ici les résultats de la seconde passation (d'où le préfixe ²) du PER 2000 aux enfants de l'enquête 2001-2006 PRS-UNADREO[37]. Le test d'Inizan a été passé simultanément, aux fins de validation externe.

Le PER 2000 comporte différentes épreuves qui ne sont pas toutes utilisées ici (notamment l'inventaire des phonèmes maîtrisés, qui présente peu d'intérêt statistique : à part ch et j, les difficultés sont rares). L'analyse de la complexité syntaxique des verbalisations en commentaire à une histoire en image est ici repérée par la série des pourcentages de phrases :

- ²%Phrases % de phrases
- ²%PhrAvEx % de phrases avec expansion
- ²%PhAvMoFo % de phrases avec monème fonctionnel
- ²%PhAvL2E % de phrases avec les deux expansions

L'épreuve de compréhension est constituée de cinq items (Qui, Combien, Où, Comment, Pourquoi). Ils sont ici additionnés en un score de 0 à 5 dans ²scorCompr.

Les tests non verbaux sont au nombre de 4 (Dessin, Sériation, Complètement, Jetons) et fournissent un score de 0 à 4 ²scoNonVer.

Les variables ²scorRythm et ²scorLogat sont les scores à l'épreuve de rythmes et à l'épreuve de logatomes.

Le test d'Inizan génère les variables suivantes :

- Inz1FigGéo Figures géométriques
- Inz2DisVis Discrimination visuelle
- Inz3MéDeDe Mémoire des dessins
- Inz4RytCop Rythmes copie
- Inz5DisPho Discrimination phonologique
- Inz6LanCom Langage compréhension
- Inz7RytRép Rythme répétition
- Inz8ArtPar Articulation parole
- Inz9LanExp Langage expression
- Inz10Cubes Cubes

[36] Se reporter à l'annexe statistique de la documentation d'Hector, et au cours de statistiques en Annexe 3.

[37] Qui a constitué le prélude expérimental aux activités Com'ens.

La collection Epreuves Inizan rassemble les variables ci-dessus, la collection ² Scores PER les variables PER énumérées auparavant.

La question de l'étalonnage

Une question peut se poser : quand on dispose d'une table d'étalonnage, comme c'est le cas pour Inizan, vaut-il mieux utiliser cet étalonnage plutôt que le score brut ? La réponse à cette question suppose une réflexion sur ce qu'est l'étalonnage. On y reviendra plus en détail dans la section C, mais il faut se souvenir que l'étalonnage

- Tient compte de l'âge à la passation du test
- Est établi sur une certaine population

On dispose aussi d'un étalonnage (Ferrand-Nespoulous) tenant compte de l'âge pour le PER. Certes, il s'agit moins d'un étalonnage que d'un seuillage en classes {normal, à surveiller, à risques}, mais ce n'est pas un obstacle dirimant à la validation externe, puisqu'on peut calculer des coefficients de corrélation par rangs entre ordinales[38].

Le problème est ailleurs : PER et Inizan n'ont pas été étalonnés sur la même population, d'une part, et d'autre part, s'agissant d'une passation simultanée des deux tests par les mêmes individus, la correction d'âge apportée par l'étalonnage est inutile. La recherche de corrélation sur les valeurs étalonnées introduirait donc inutilement un élément perturbateur : les différences de population, voire de méthode d'étalonnage (celle d'Inizan a varié dans le temps[39]).

Résultats

Les deux collections étant installées, on clique le bouton calcul de la matrice de statistiques.

On obtient le tableau suivant :

Collection Epreuves Inizan × collection ² scores PER

Matrice des coefficients de corrélation r (Bravais-Pearson)

	²%Phrases	²%PhrAvEx	²%PhAvMoFo	²%PhAvL2E	²scorCompr	²scoNonVer	²scorRythm	²scorLogat
Inz1FigGéo						0.367 ***	0.231 **	0.272 ***
Inz2DisVis						0.314 ***	0.252 **	
Inz3MéDeDe						0.267 **		
Inz4RytCop				0.235 **		0.407 ***	0.236 **	0.228 **
Inz5DisPho						0.297 ***	0.230 **	0.314 ***
Inz6LanCom								0.259 **
Inz7RytRép							0.259 **	0.300 ***
Inz8ArtPar						0.223 **	0.231 **	0.606 ***
Inz9LanExp		0.223 **	0.264 **	0.259 **		0.281 ***		0.373 ***
Inz10Cubes						0.422 ***	0.248 **	0.238 **

A l'intersection de la ligne d'une variable de l'une des collections (ici Inizan) et de la colonne d'une variable de l'autre collection (ici PER, mais on aurait pu aussi dresser le tableau dans l'autre sens), on trouve le coefficient de corrélation de bravais-Pearson. Certaines cases semblent vides : c'est qu'on a coché la case p<.05, et qu'en conséquence seuls sont affichés les coefficients significatifs au moins au seuil de .05.

Les corrélations significatives à .05 portent deux astérisques, les corrélations significatives à .01, trois. Le tableau n'est pas symétrique, puisqu'il s'agit du croisement de deux

[38] Voir le type des variables dans le cours de statistiques.

[39] D'abord une normalisation, puis un quantilage.

collections et non d'une collection avec elle-même, comme c'était le cas dans l'étude des parentés entre items en validation interne. On notera que la réalisation manuelle de ce tableau aurait demandé autant de croisements de deux variables qu'il y a de cases à l'intérieur du tableau, soit 80 croisements.

La significativité est une chose, la force de la corrélation en est une autre. Une corrélation peut être très significative sans être forte, surtout quand l'effectif est important (ici, plus de 400 sujets), puisque la significativité dépend aussi de l'effectif.

Selon les propositions de Piéron (1969), les corrélations observées dans la matrice vont de faible à moyenne, à l'exception sans doute de la corrélation entre Inizan Articulation parole et PER score Logatomes, qui, avec .606, est assez forte.

Interprétation et variante

Comment interpréter un tel tableau ? Certaines corrélations sont plus fortes que d'autres, mais aussi certaines épreuves d'Inizan n'ont logiquement rien à voir avec certaines épreuves du PER, et même si certains domaines se recoupent, nous avons écrit plus haut que l'on ne recherchait pas une correspondance absolue, ce qui serait absurde, mais une certaine consistance, pour laquelle les critères sont évidemment moins exigeants.

Il paraît difficile ici au statisticien de poursuivre, et l'interprétation du tableau, qui met en jeu la connaissance fine du contenu des tests, est du ressort du chercheur en Orthophonie.

Tout au plus ajoutera-t-on que si le test de référence était supposé couvrir tout ce que le test à valider est supposé aborder, fût-ce en raccourci (ce qui ne semble pas être le cas ici), alors le fait de trouver des épreuves du test de référence qui ne sont corrélées avec aucune épreuve du test à valider laisserait à penser que celui-ci comporte des lacunes.

Si l'on a des raisons de douter de la qualité métrique des variables utilisées, par exemple parce qu'elles sont très asymétriques, on remplacera le coefficient de corrélation de Bravais-Pearson, qui n'est en principe applicable qu'aux distributions approximativement normales, par le coefficient de corrélation par rangs rhô de Spearman, qui n'a pas le même niveau d'exigence.

Cela se fait en sélectionnant les tests souhaités par les cases à cocher dans la Matrice de statistiques.

Matrice des coefficients de corrélation par rangs rhô de Spearman

	²%Phrases	²%PhrAvEx	²%PhAvMoFo	²%PhAvL2E	²scorCompr	²scoNonVer	²scorRythm	²scorLogat
Inz1FigGéo						0.359 ***		0.266 **
Inz2DisVis						0.249 **		
Inz3MéDeDe						0.233 **		
Inz4RytCop				0.221 **		0.392 ***	0.212 **	0.220 **
Inz5DisPho						0.309 ***		0.336 ***
Inz6LanCom								
Inz7RytRép							0.277 ***	0.341 ***
Inz8ArtPar						0.229 **		0.593 ***
Inz9LanExp			0.249 **	0.281 ***		0.303 ***		0.292 ***
Inz10Cubes						0.400 ***		

Les résultats sont assez peu différents.

Cas de variables non-numériques

Un autre exemple, emprunté au travail de deux autres étudiantes[40] en quatrième année d'Orthophonie, offre l'occasion de considérer la validation externe dans le cas où les variables ne sont pas numériques.

On souhaite valider à l'externe un test Odédys par référence à la notion de « seuil du savoir lire » d'après le test d'Inizan.

(Classement Odédys partiel)

	effectifs	%/Total	% cumulés
à risque	9	12.50%	12.50%
à surv.	12	16.67%	29.17%
satisf.	51	70.83%	100.00%
Total	72	100.00%	

L'Odédys est dit ici partiel parce que, suite à la démarche de validation interne, certains éléments ont été écartés.

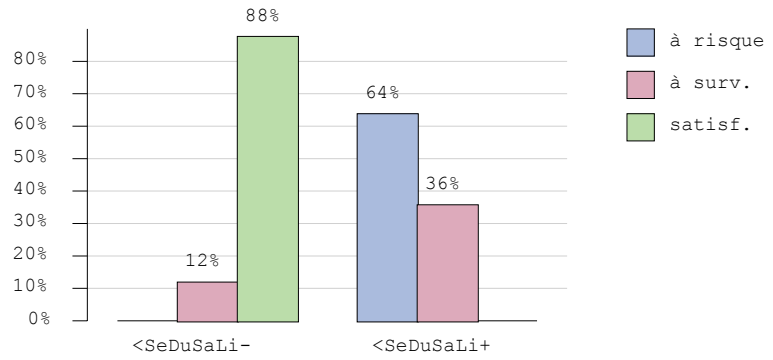
Le découpage en trois catégories {à risque, à surveiller, satisfaisant} a été obtenu par les procédés de seuillage présentés à la section suivante.

En croisant ce classement[41] avec le seuil du savoir lire d'Inizan, on obtient le tableau suivant[42] :

(< seuil du savoir lire) × (Classement Odédys partiel)

N %L +	à risque	à surv.	satisf.	S/LIGNE :
<SeDuSaLi-	---	7 12% --	51 88% +++	58 100%
<SeDuSaLi+	9 64% +++	5 36% ++	---	14 100%
S/COLONNE:	9 12%	12 17%	51 71%	72 100%

rhô (Spearman) = -0.821 , s. à .01



[40] Remerciements à SEVERINE BARON ET SYLVIE LECONTE, (2007), *Complément de validation du DPL3. Quelle prédictivité par rapport aux compétences du lecteur/scripteur débutant*. Mémoire présenté pour l'obtention du certificat de Capacité d'Orthophoniste, sous la direction de Françoise COQUET. Institut d'Orthophonie Gabriel Decroix, Université de Lille II.

[41] Il s'agit d'un simple croisement classique de deux variables dans l'onglet Tris de la page Traitements d'Hector.

[42] Attention à un possible problème d'interprétation : <SeDuSaLi- signifie « non inférieur au seuil du savoir lire ».

Comme le classement est une variable ordinale, le rhô de Spearman est applicable : il est ici très élevé (et négatif pour des raisons sémantiques : « <SeDuSaLi+ » signifie « inférieur au seuil du savoir lire).

On note que si tous les enfants « à risque » au sens d'Odedys sont également repérés par Inizan, et que tous les enfants « satisfaisants » au sens d'Odedys le sont aussi pour Inizan, les enfants « à surveiller » au sens d'Odédys constituent à juste titre une catégorie ambiguë au sens d'Inizan, puisqu'ils se répartissent à peu près également de part et d'autre du seuil.

On peut considérer qu'Odédys est, pour cet échantillon, clairement validé par Inizan, le flottement inévitable de la classe intermédiaire étant affaire de sensibilité des tests.

Qualités prédictives d'un test

On poursuit avec le même ensemble de données. Outre la validation interne et externe d'Odédys, le projet comportait la vérification des qualités prédictives de la grille d'observation DPL3, passée par les mêmes enfants quelques années plus tôt.

Les qualités prédictives d'un test concernent sa capacité à anticiper un phénomène ultérieur, non encore mesurable au moment où le test est passé. Le modèle sous-jacent est que les processus (apprentissage, maturation) pouvant influer sur les capacités étudiées ont eu la même influence sur tous les sujets concernés, ceci constituant la version appropriée du « toutes choses étant égales par ailleurs ».

Il est clair que dans les conditions écologiques de l'apprentissage de la langue maternelle ce contrôle des facteurs est illusoire. On s'attendra donc à des prédictions relativement approximatives, l'essentiel étant surtout de pouvoir lever précocement un signal d'alarme.

Ici, le contrôle au temps t_1 du DPL3 passé au temps t_0 est effectué sur la double base de l'Odédys et du seuil du savoir lire d'Inizan.

Le DPL3 fournit un classement selon les trois niveaux classiques :

(Classement DPL3)

	effectifs	%/Total	% cumulés
a risque	4	5.56%	5.56%
a surv	6	8.33%	13.89%
satisf	62	86.11%	100.00%
Total	72	100.00%	

La prédictivité s'étudie généralement sur la base d'un tableau 2 x 2, qui permet l'analyse des risques d'erreur α et β. On réduit donc la variable « Classement DPL3 » à deux positions dans une variable « dépistés DPL3 », les « dépistés » regroupant les « à surveiller » et les « à risques ». Les faibles effectifs concernés auraient de toutes façons suggéré ce regroupement. De la même manière, le classement Odédys est ramené à deux positions :

(dépistés DPL3) × (dépisté Odédys partiel)

N %L +	DépOdéPar-	DépOdéPar+	S/LIGNE :
DépisDpl3-	49 79% +++	13 21% ---	62 100%
DépisDpl3+	2 20% ---	8 80% +++	10 100%
S/COLONNE:	51 71%	21 29%	72 100%

Khi2 = 12.87 pour 1 d.d.l. avec 1 correction(s) de Yates, s. à .01

Les DépisDpl3+ sont dépistés par Dpl3 comme pouvant à l'avenir poser problème. Les DépOdéPar+ sont dépistés par Odédys comme posant actuellement problème. La question qu'on se pose est « rétrospectivement, DPL3 aurait-il permis de s'attendre aux futures résultats Odédys ? » [43].

Conformément à la tradition médicale, « positif » désigne les individus pour lesquels le test indique qu'il y a un problème.

Les 2 individus que Dpl3 dépiste, mais qui sont négatifs pour Odédys (pas de problèmes) sont appelés *faux positifs*, et tant mieux pour eux ; les 13 individus négatifs pour Dpl3, mais qu'Odédys dénonce sont appelés *faux négatifs*, et c'est plus ennuyeux, parce que s'ils avaient été dépistés il aurait peut être été possible de remédier.

Le khi2 significatif à .01 nous indique que les données observées s'éloignent significativement du modèle de l'indépendance entre les deux variables, *et qu'il est donc légitime de commenter le tableau*.

Un certain nombre de paramètres peuvent être calculés sur la base de ce tableau[44] :

- La *sensibilité* est, parmi les positifs dans la mesure actuelle, la proportion de ceux qui étaient positifs dans la mesure prédictive : ici 8/21 soit 38%
- La *spécificité* est, parmi les négatifs de la mesure actuelle, la proportion de ceux qui étaient négatifs dans la mesure prédictive : ici 49/51 soit 96%
- La *prédictivité* est le taux de prédictions qui se sont avérées justes : positifs les deux fois ou négatifs les deux fois : ici (49+8)/72 soit 79%

Qu'en conclure ? A l'aune de l'Odédys, le Dpl3 est correctement prédictif avec près de 80% de prédictions justes. Il est très spécifique, ce qui signifie qu'il est bien approprié à prédire le phénomène que mesure Odédys (et non un autre), mais sa faiblesse réside dans sa relativement faible sensibilité, liée au nombre élevé de faux négatifs.

Dans une prédiction parfaite, il n'y aurait ni faux positifs, ni faux négatifs, et la prédictibilité, comme la sensibilité et la spécificité, serait égale à 100%. Mais, en deçà, qu'est-ce qui est acceptable ? Nous ne fournirons pas de seuils d'acceptabilité pour ces paramètres – peut-être la littérature spécialisée en donne-t-elle – et nous bornerons à indiquer que ces éléments servent surtout à enclencher une réflexion analytique sur les épreuves utilisées, et éventuellement à comparer entre elles les qualités prédictives de plusieurs tests. Si l'on croise maintenant les résultats du DPL3 avec le seuil du savoir lire au sens d'Inizan, on obtient le tableau suivant :

(dépistés DPL3) × (< seuil du savoir lire)

N %L+	<SeDuSaLi-	<SeDuSaLi+	S/LIGNE :
DépisDpl3-	51 82%	11 18%	62 100%
DépisDpl3+	7 70%	3 30%	10 100%
S/COLONNE:	58 81%	14 19%	72 100%

Khi2 = 0.41 pour 1 d.d.l. avec 1 correction(s) de Yates, n.s.

[43] On remarquera que la confrontation prédictive d'un test à un autre test pose un problème épistémologique, puisqu'on dispose d'indicateurs tous deux « obliques » d'une compétence non observable globalement en tant que telle : les différences observées peuvent être en partie imputables aux biais propres à ces tests. Il eût mieux valu être en possession d'une variable du type « a effectivement appris à lire », mais on doit souvent faire avec ce qu'on a.

[44] Dans les versions d'Hector postérieures à Juillet 2008, le calcul de ces taux est effectué automatiquement pour les croisements de logiques, à condition d'avoir coché la case [Pr] dans le panneau des options.

Il faut ici résister à la tentation de calculer les paramètres comme ci-avant, car le khi2 non significatif nous indique que les données observées ne s'écartent pas significativement d'une répartition au hasard, et *qu'il est donc illégitime de commenter le tableau*.

En conclusion, dans cet échantillon, Dpl3 est prédictif pour Odédys, mais non pour Inizan.

Fidélité, mesures avant/après

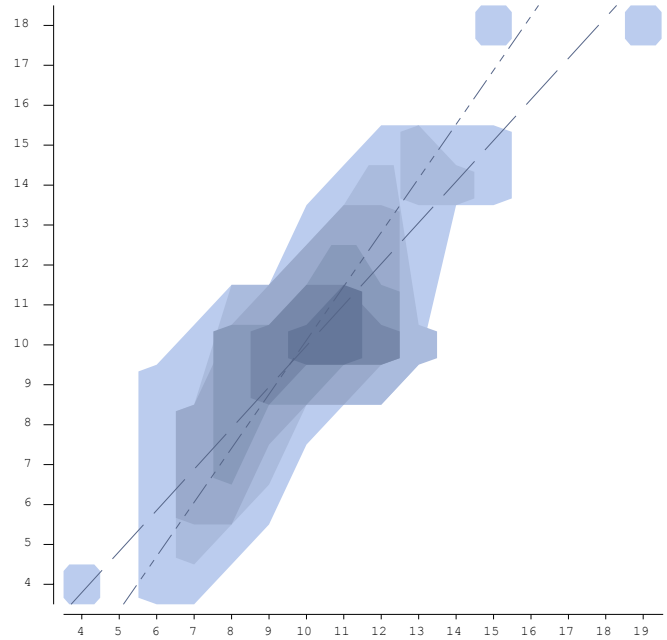
Fidélité, le souci de la docimologie

La *fidélité* d'une épreuve est sa tendance a donner en t_1 les *mêmes résultats* qu'en t_0 sur les *mêmes individus*. Le modèle sous-jacent est celui d'un instrument de mesure qui ne se dérègle pas.

La docimologie moderne est née de recherches sur la fidélité de la notation aux examens[45]. La question pouvait être résumée ainsi : « un professeur a qui l'on redonne à corriger, après un certain temps, les mêmes copies anonymées, va-t-il redonner les mêmes notes ? ». En matière de tests, la question devient : « Si on refait passer le même test à quelqu'un après un certain temps, obtiendra-t-on le même résultat ? »

La mesure usuelle de la fidélité est un simple coefficient de corrélation de Bravais-Pearson :

```
(note 1) × (note 2)
r (Bravais-Pearson) = 0.872 , s. à .01
```



Les notes 1 et 2 sont supposées avoir été attribuées, à quelques mois d'intervalles, par le même correcteur, à un même paquet de copies anonymées[46]. Le nuage de densité ci-dessus est d'autant plus sombre que de nombreux sujets se concentrent dans telle ou telle zone.

[45] Travaux de Laugier et Weinberg, dans les années 1930, sur les copies du baccalauréat. Ces travaux portaient également sur la correction multiple, qui relèverait plutôt de la cohérence et de la fiabilité.

[46] Les données de cet exemple ont été fabriquées exprès pour illustrer la démarche. Il ne s'agit pas de données issues d'une véritable expérimentation.

Si la concordance était parfaite, toutes les notes s'aligneraient exactement sur une diagonale et le coefficient de corrélation serait égal à 1.

Ce n'est pas tout à fait le cas, il est de .872, ce qui dans des conditions réelles constituerait une *fidélité* tout à fait remarquable[47].

Pourtant, si on étudie la différence, pour chaque individu, la différence entre sa note lors de la première correction et lors de la seconde, on observe des écarts importants :

(diff 1 2)

	effectifs	%/Total	% cumulés
-3	2	3.28%	3.28%
-2	10	16.39%	19.67%
-1	12	19.67%	39.34%
0	13	21.31%	60.66%
1	16	26.23%	86.89%
2	4	6.56%	93.44%
3	4	6.56%	100.00%
Total	61	100.00%	

L'écart va de 3 points en moins à 3 points en plus. Pourtant, ce correcteur est plutôt fidèle à sa propre notation !

Difficultés de la mesure de fidélité pour les tests

Une difficulté fondamentale concernant la fidélité dans les tests orthophoniques est que la fidélité est supposer être mesurée, *rien n'ayant changé par ailleurs*. Avec des enfants en plein apprentissage et dans un métier voué à la rééducation, les conditions ne sont pratiquement jamais remplies.

Le cas limite où cela serait possible concernerait la mesure de compétences (linguistiques ou autres) chez des adultes en bonne santé. Un obstacle sournois, rencontré par les constructeurs de tests d'intelligence (et leurs petits cousins les recruteurs), est la *test-wiseness*, autrement dit le fait que passer un test provoque un effet latéral d'apprentissage du test[48]. C'est très embarrassant pour un instrument supposé mesurer une dimension permanente de l'individu !

La solution est recherchée du côté de l'usage de *formes parallèles* du même test, c'est-à-dire d'une variante avec d'autres items, mais qui mesure la même chose en contournant l'apprentissage lié à la première passation. Bien sûr, il faut avoir préalablement prouvé, sur un autre échantillon (inutilisable ensuite) et par une variante de la validation interne, que les deux formes sont effectivement parallèles …

Avant et après, le rêve des pédagogues

La mesure de fidélité, pour laquelle rien n'est supposé changer, ne doit pas être confondue avec les mesures avant/après, où l'on cherche précisément à vérifier si quelque chose a changé.

C'est le cas typique d'une situation d'apprentissage, où l'on fait passer le même test (ou une forme parallèle, ou en espérant qu'assez de temps s'est écoulé pour qu'on oublie) à des individus *avant* l'apprentissage, puis *après* l'apprentissage.

Logiquement, si l'apprentissage est efficace, les résultats devraient être supérieurs après.

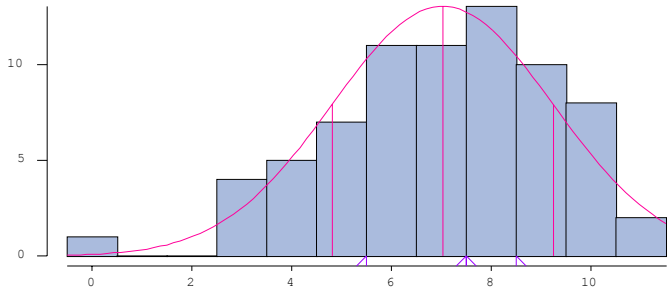
---

[47] Le maximum observé par Laugier et Weinberg était de .81, mais les valeurs courantes tournaient autour de .60.

[48] Il est extrêmement difficile d'empêcher quelqu'un d'apprendre, tous les enseignants vous le diront. Apprendre quoi, c'est autre chose.

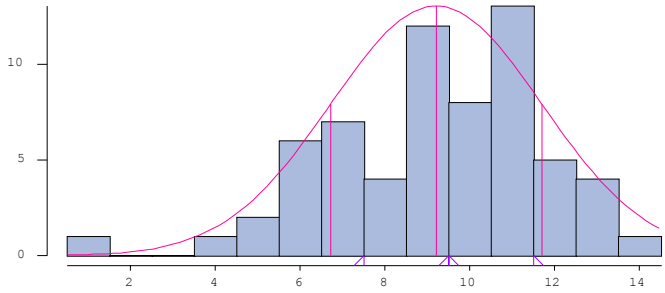
Une première mesure avant donne les résultats suivants[49] :

```
(score avant)
Valeur modale : 8 (n=13)
Médiane entre 7 & 8
Moyenne 7.03, écart-type 2.22
```



Après l'apprentissage, les résultats ont l'allure suivante :

```
(score après)
Valeur modale : 11 (n=13)
Médiane entre 9 & 10
Moyenne 9.20, écart-type 2.49
```

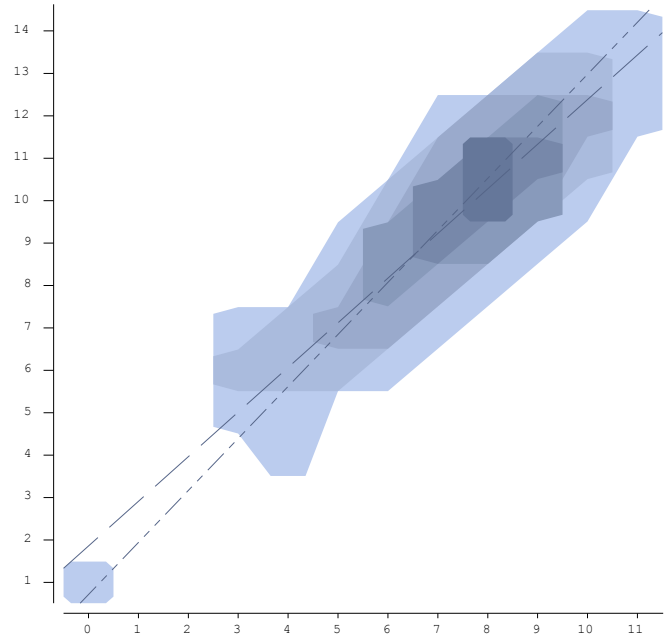


Les choses semblent s'être améliorées : la moyenne est passée de 7.03 à 9.20.

Pour en avoir le cœur net, on croise les deux mesures :

```
r (Bravais-Pearson) = 0.926 , s. à .01
dm = 2.19 ± 0.96 ; |t| (Student) = 18.278 , s. à .01
```

[49] Là aussi, il s'agit de données fabriquées tout exprès.



Certes, elles sont fortement corrélées (si tous les élèves ont progressé, la hiérarchie entre eux a peu évolué), mais le test statistique le plus important dans ce cas est le |t| de Student sur échantillons appareillés (que l'on sélectionne dans le panneau des options auquel on accède par le bouton du même nom), qui teste l'hypothèse selon laquelle les différences de scores entre avant et après sont différentes de zéro.

Ici, la différence entre les scores, très significative, est de 2,19 points.

Dans une perspective expérimentaliste que nous ne détaillerons pas parce qu'elle s'écarte trop de ce document (consacré à la validation des tests), on pourra aussi disposer d'un groupe témoin auquel aucun apprentissage n'a été proposé, et qui ne devrait donc pas progresser en principe ; on créera une variable codant la différence des scores avant/après pour chaque individu, et on croisera cette différence avec l'appartenance ou non au groupe témoin, en utilisant le F de Snédécor-Fisher (Anova) pour déterminer si le groupe expérimental a d'avantage progressé.

Un exemple synthétique

Dans une démarche complète d'étude de fidélité et de prédictivité, les croisements utilisés sont très nombreux, et on propose ici une présentation synthétique des résultats :

- la colonne rBP indique la corrélation de Bravais-Pearson et son seuil de probabilité s'il est significatif
- la colonne ρ indique la corrélation de Spearman et son seuil de probabilité s'il est significatif (cette corrélation sur les rangs est souvent plus élevée que rBP quand les distributions ne sont pas approximativement normales)
- la colonne χ^2 fournit la valeur du khi-carré sur le croisement dichotomique et son seuil de probabilité s'il est significatif.
- Les valeurs de sensibilité, de spécificité et de prédictivité ne sont fournies que si le χ^2 est significatif. Pour leur signification détaillée, se reporter à la rubrique *Qualités prédictives*.

Les données sont issues d'une passation des mêmes épreuves à 20 et 27 mois[50]. Les corrélations sont calculées sur des scores bruts, et les variables logiques sont des dichotomies non pas sur la médiane, mais sur le premier quartile, de manière à opposer un quart de « faibles » à trois quarts de « normaux ».

Les dichotomies au seuil du premier quartile ne sont souvent pas définies ici parce que la variable d'origine n'a pas assez de valeurs distinctes. Elles n'ont jamais de χ^2 significatif, on ne les a donc pas fait figurer dans ce tableau.

Epreuve	rBP	ρ	χ^2	sensibilité	spécificité	prédictivité
Désignation parties du corps	.158	.090	0,73			
Désignation objets	.221 .10	.200	0,02			
Réalisation ordre simple/ CMS	.348 .01	.346 .01	4,71 .05	87%	38%	63%
Dénomination objets	.419 .01	.452 .01	6,05 .05	68%	64%	66%
Phrases à deux mots/ PMS	.343 .01	.295 .05	3,39 .10	82%	40%	68%
Praxies bucco-faciales	-.086	-.109	0,64			
Objets spontanés	.406 .01	.404 .01	7,12 .01	69%	67%	68%
Permanence objet	.234 .10	.234 .10	1,31			
Attention conjointe	0	0	1,31			

[50] Données empruntées à Perrine CUCUEL et Anne-Claire MOREL (2008), *Protocoles d'évaluation des fonctions de communication et de langage chez des enfants de 20 puis 27 mois*. Mémoire présenté pour l'obtention du certificat de Capacité d'Orthophoniste, sous la direction de Françoise COQUET et Valérie DELPORTE. Institut d'Orthophonie Gabriel Decroix, Université de Lille II.

Section C : étalonnage et seuils critiques

L'étalonnage est l'opération qui consiste à étudier la distribution d'un ou plusieurs scores sur une grande population, de manière à construire des échelles ordinales permettant le repérage rapide de la position d'un individu au regard de la population de référence. Plusieurs techniques sont envisageables, selon les approches paramétriques ou non-paramétriques ; parmi ces dernières, les deux plus utilisées sont la normalisation et le quantilage.

L'approche paramétrique

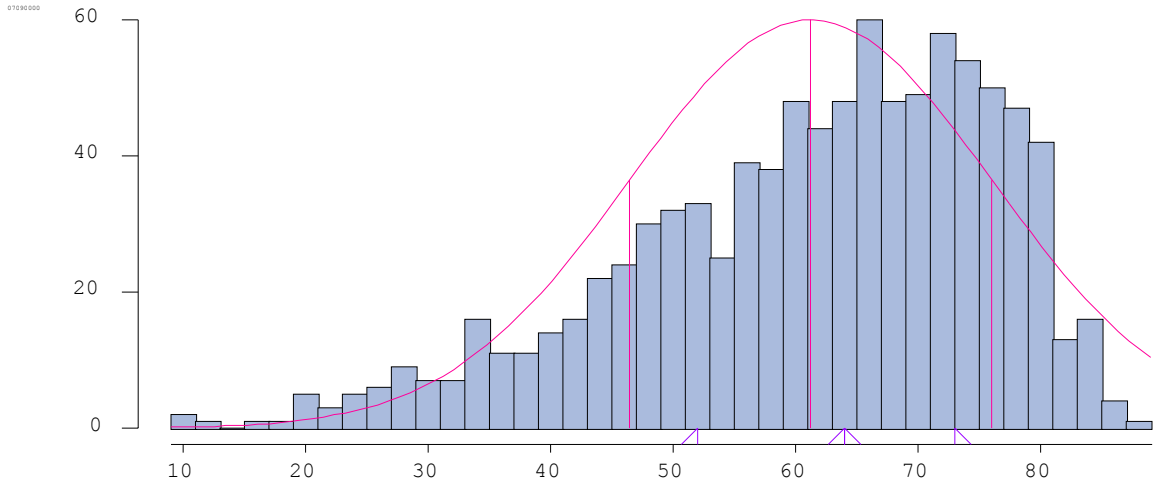
Elle consiste à s'appuyer sur les *paramètres* d'une distribution, à savoir essentiellement sa *moyenne* et son *écart-type*. Classiquement, elle sert à distinguer des individus *atypiques faibles*, qui obtiennent un score inférieur à la moyenne, moins un[51] écart-type, symétriquement des *atypiques forts*, qui obtiennent un score supérieur à la moyenne, plus un écart-type, et entre deux les *typiques*.

Cette approche est d'une grande simplicité, puisqu'il suffit de fournir la moyenne et l'écart-type de la distribution du score à une épreuve pour que celle-ci soit réputée étalonnée.

Elle est pour cette raison très utilisée, et passe pour *la* méthode naturelle. Elle a cependant un grave défaut : elle suppose que la distribution est approximativement normale, faute de quoi la moyenne et l'écart-type ne peuvent pas être considérés comme des résumés utilisables des caractéristiques de la distribution.

Ainsi dans l'exemple suivant[52] :

```
(total maths)
Classe modale : [66,68[ (n=60)
Médiane entre 64 et 65
Moyenne 61.74, écart-type 14.81
H(normalité) rejetée à .0000 ; H(symétrie) rejetée à .0001
```



Cette distribution n'est ni normale, ni même grossièrement symétrique. Il est tout à fait impossible d'en faire usage dans une approche paramétrique, et si on cherche à la corréler avec une autre, on sera bien avisé d'utiliser la corrélation par rangs rhô de Spearman.

[51] Parfois moins deux écarts-types, ou un écart-type et demi.

[52] Il s'agit du score total en mathématiques aux évaluations CE2 en 2004 dans deux circonscriptions.

Il est donc recommandé, avant tout emploi de l'approche paramétrique, de s'assurer qu'elle est appropriée.

Normalisation

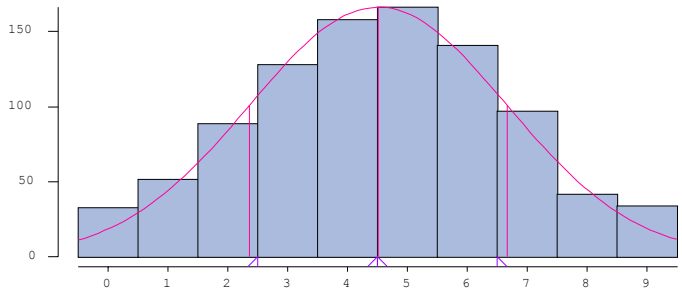
Reprenons l'exemple précédent, celui d'une variable peu normale, asymétrique et irrégulière : le score brut d'un individu ne renseigne pas sur son positionnement par rapport à l'ensemble.

On crée une nouvelle variable par la formule de calcul suivante[53] :

```
# normalisation_math
: NORM( total_Math 10 ) ;
```

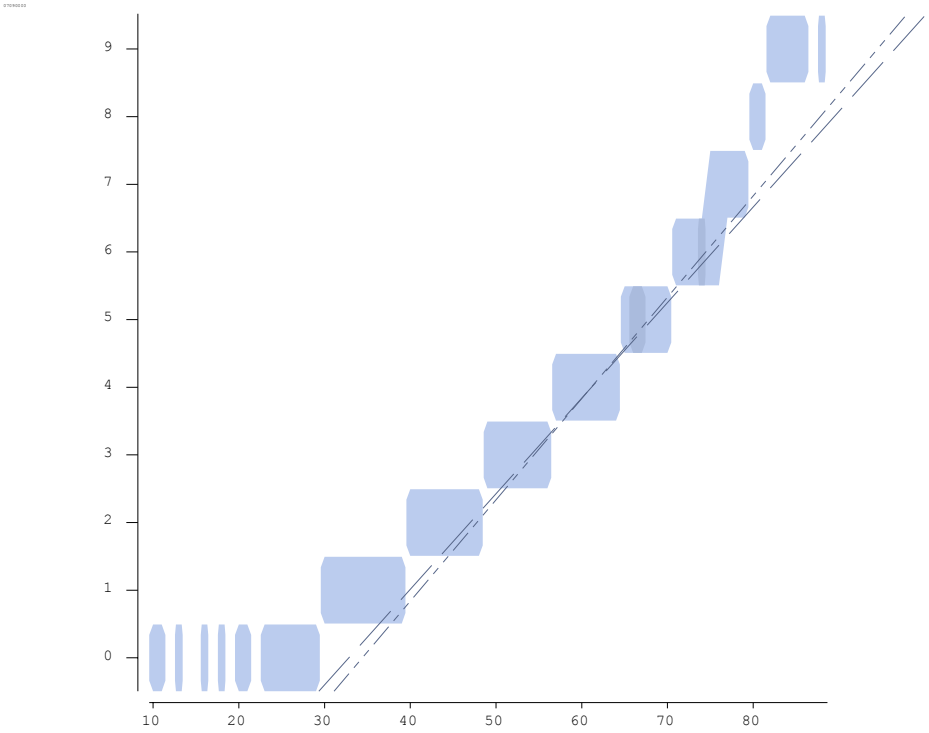
Cette nouvelle variable prend ses valeurs entre 0 et 9, et les effectifs de ses cases successives sont approximativement ceux que montrerait une distribution normale sur 10 positions.

```
(normalisation math)
```



L'ajustement est approximatif parce que les sujets ne peuvent se partager entre plusieurs cases. Si l'on croise cette nouvelle variable avec sa variable-source, on obtient ceci :

```
(total maths) x (normalisation math)
```



[53] La syntaxe du langage des formules d'Hector est décrite dans le manuel Hector 2 variables formulées.pdf.

On voit comment plusieurs valeurs d'origine se regroupent pour « remplir » une case de la nouvelle variable.

La table de conversion résultant d'une telle transformation ressemblerait à ceci :

Scores bruts	classe
< 30	0
30 à 39	1
40 à 48	2
49 à 56	3
57 à 64	4
65 à 70	5
71 à 75	6
76 à 79	7
80, 81	8
> 81	9

A un score brut de 66 correspondrait la classe 5, à un score brut de 78, la classe 7.

Attention ! Cette nouvelle variable est initialement créée comme numérique, mais en fait elle n'est qu'ordinale : son seul point commun avec la variable d'origine, c'est *l'ordre* des individus. De ce fait, l'addition de « scores normalisés », qui ne sont en fait que des « numéros de classe normalisées », est épistémologiquement discutable, quoique très répandue. Ainsi, additionner un score normalisé dans une épreuve A avec un score normalisé dans une épreuve B n'a aucun sens.

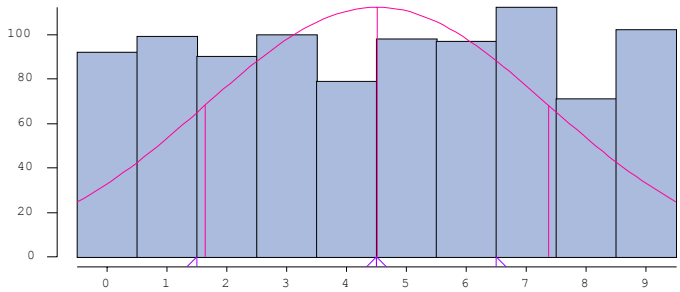
Quantilage

Cette transformation relève du même principe que la normalisation, sauf qu'au lieu de chercher à obtenir des classes dont les effectifs sont approximativement proportionnels à ceux d'une distribution normale, on cherche à obtenir des classes aux effectifs à peu près égaux (distribution plate).

La formule est très similaire à la précédente :

```
# quantiles_math
: QUANT( total_Math 10 ) ;
```

Les classes sont approximativement équiprobables :



La table de conversion sera évidemment différente :

- 35 était en 1, il sera en 0
- 44 était en 2, il sera en 1
- 66 est toujours en 5 (les systèmes coïncident au milieu), mais 78 est maintenant en classe 8

Scores bruts	classe
< 40	0
41 à 49	1
50 à 55	2
56 à 60	3
61 à 64	4
65 à 67	5
68 à 71	6
72 à 75	7
76 à 78	8
> 79	9

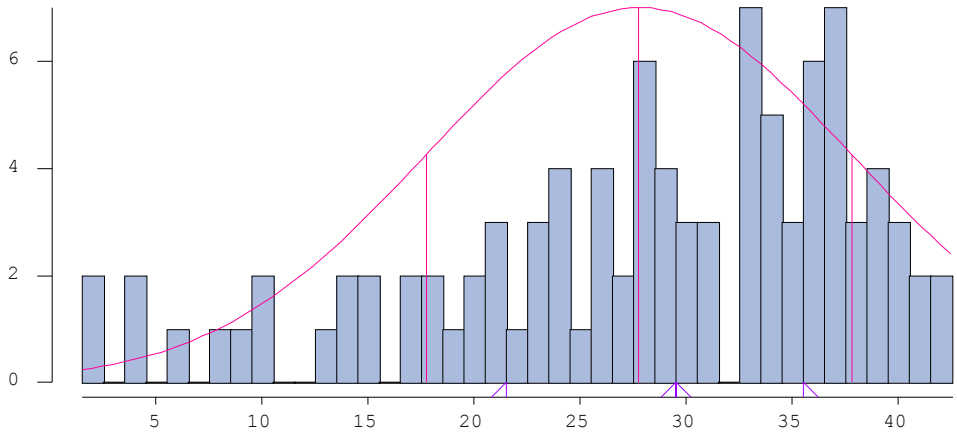
Quel système est préférable ? C'est affaire de goût : la normalisation isole mieux les cas extrêmes, le quantilage utilise plus largement l'éventail des classes. Dans les deux cas, le nombre de cases est à la discrétion de l'usager, mais 10 cases est assez courant, de 0 à 9 ou de 1 à 10, en dépit du risque de confusion avec des notes scolaires.

Attention, dans les deux cas, concernant des compétences à développement rapide, comme c'est souvent le cas en matière d'apprentissage du langage, il convient d'*étalonner séparément* par tranches d'âge, voire par sexe (comme dans le quintilage du test du Bonhomme).

Les seuils critiques

La fixation de seuils critiques, ou la création de classements à trois niveaux {normal, à surveiller, à risques} participe du même esprit que l'étalonnage, de manière simplifiée. Traditionnellement, comme on l'a vu plus haut, on rencontre dans les distributions de scores la notion *d'atypiques faibles*, qui désigne les individus dont la mesure se situe en dessous de la moyenne moins un écart-type, voire d'atypiques très faibles au delà de deux écarts-types.

Cet appel aux paramètres de résumé central et de dispersion d'une distribution numérique, ou seuillage paramétrique, rencontre concrètement des difficultés, notamment quand les distributions ne sont pas régulières[54] :




```
Valeur modale : 33 (n=7)
Médiane entre 29 & 30
```

[54] Données issues, comme plus haut, du mémoire Deprey-Renard 2007

Moyenne 27.75, écart-type 10.05

Cette distribution est très asymétrique, avec une queue dispersée à gauche.

La coupure à un écart-type en dessous de la moyenne, entre 16 et 17, rassemble trop de sujets pour les définir « à risques », et sans doute pas assez pour une catégorie « à surveiller ». D'autre part, le creux de la courbe entre 10 et 13 semble indiquer la possibilité d'une coupure « naturelle » : plus l'effectif est rare au point de coupure, plus le risque d'erreur est faible.

On préfère donc suggérer deux coupures non paramétriques :

L'une entre 11 et 13, dans le creux de la courbe

L'autre au premier quartile (le petit triangle bleu en bas de l'histogramme, entre 21 et 22.

(score total)

	effectifs	%/Total	% cumulés
2	2	2.06%	2.06%
4	2	2.06%	4.12%
6	1	1.03%	5.15%
8	1	1.03%	6.19%
9	1	1.03%	7.22%
10	2	2.06%	9.28%
			coupure
13	1	1.03%	10.31%
14	2	2.06%	12.37%
15	2	2.06%	14.43%
17	2	2.06%	16.49%
18	2	2.06%	18.56%
19	1	1.03%	19.59%
20	2	2.06%	21.65%
21	3	3.09%	24.74%
			coupure
22	1	1.03%	25.77%
23	3	3.09%	28.87%
24	4	4.12%	32.99%
25	1	1.03%	34.02%
26	4	4.12%	38.14%
27	2	2.06%	40.21%
28	6	6.19%	46.39%
29	4	4.12%	50.52%
30	3	3.09%	53.61%
31	3	3.09%	56.70%
33	7	7.22%	63.92%
34	5	5.15%	69.07%
35	3	3.09%	72.16%
36	6	6.19%	78.35%
37	7	7.22%	85.57%
38	3	3.09%	88.66%
39	4	4.12%	92.78%
40	3	3.09%	95.88%
41	2	2.06%	97.94%
42	2	2.06%	100.00%
Total	97	100.00%	

Les coupures sont reportées dans le tableau ci-dessus, colonne des pourcentages cumulés.
La table de seuillage résultante s'établit donc ainsi :

score	classe	
< 12	en difficulté	9%
12 à 21	fragile	16%
> 21	satisfaisant	75%

Si l'on tient absolument à se rapprocher d'une référence paramétrique, on notera que, par coïncidence, les pourcentages obtenus correspondent approximativement, dans la Loi Normale à des coupures réalisées à moyenne –2/3 d'écart-type et moyenne – 4/3 d'écart-type, valeurs qui semblent en l'occurrence plus intéressantes que –1 et –2.

Chrono-cohérence et crise d'effectifs

Quand les populations-cibles sont stratifiées par classes d'âge et que le performances sont réputées varier avec l'âge, il faut en principe fournir un étalonnage distinct pour chaque classe d'âge. Le problème est évidemment qu'en découpant l'échantillon étudié en sous-échantillons, on arrive à des effectifs par sous-ensemble si faibles qu'ils défient la statistique, et empêchent notamment les vastes étalonnages par normalisation ou quantilage avec une dizaine de cas distincts.

Dans l'exemple ci-dessous, tiré du mémoire déjà cité Buret-Nicli, les effectifs disponibles ne permettent rien de plus fouillé qu'un découpage en quartiles, c'est-à-dire en quatre fractions d'effectif approximativement égaux, et ce en dépit du fait qu'il avait déjà fallu passer de six classes d'âge de 6 mois à trois classes d'âge d'un an pour causes de différences non-significatives.

Même ainsi, le faible nombre de valeurs différentes et la manière peu équilibrée dont elles s'agencent ne permettent pas toujours de distinguer des quartiles voisins. On peut cependant fournir encore un instrument utilisable en acceptant ce fait, et en présentant le résultat de la manière suivante :

Classe d'âge	Q1	Q2	Q3	Q4
3a3-4a3	0 à 2	3 à 8	9 à 11	12
4a3-5a3	0 à 7	8 à 11	12	
5a3-6a3	0 à 11	12		

Pour la tranche d'âge 3a3-4a3, on a bien quatre quartiles nettement distincts.

A 4a3-5a3, il n'est plus possible de distinguer Q3 de Q4, parce qu'environ la moitié du sous-échantillon obtient le score de 12.

A 5a3-6a3, le premier quartile recouvre tous les scores de 0 à 11, parce que les trois-quarts du sous-échantillon de cet âge obtient le score de 12.

Cet étalonnage, pour grossier qu'il soit, reste parfaitement utilisable. Simplement, il caractérise une épreuve qui est d'autant plus classante que les enfants sont jeunes, et qui ne l'est pratiquement plus pour les plus âgés. L'impossibilité de faire mieux tient non seulement à la faiblesse des effectifs, mais aussi au faible nombre de valeurs distinctes.

Section D : Miscellanea

Cette section rassemble des questions qui ne se rattachent pas clairement à l'organisation générale du document, mais qui peuvent se rencontrer en situation d'analyse d'épreuves. Elle est peut-être destinée à s'agrandir au fil des versions successives.

Probabilité des réponses et scores significatifs

Le problème dont il est question ici s'expose plus facilement à travers un exemple[55] :

Il s'agit d'une épreuve de discrimination phonétique entre paires de logatomes proches, passée à des enfants de la moyenne section de maternelle au CM2. La série de logatomes est différente pour la maternelle et pour le primaire. On ne s'intéresse ici qu'à la maternelle.

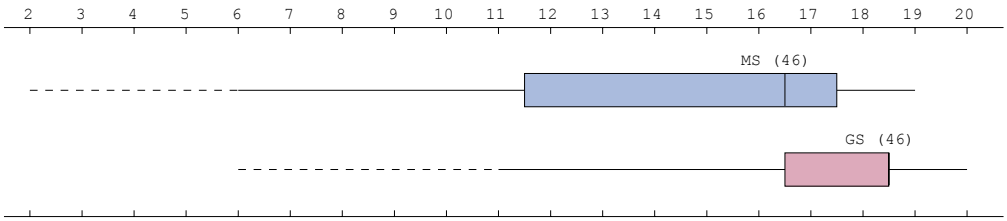
Deux modalités de passation sont successivement appliquées, l'une dite Normal, l'autre dite Rapide. Pour chaque paire, l'enfant doit répondre « pareil », « pas pareil » ou « je ne sais pas ». Les éléments d'une paire de logatomes sont, soit identiques, soit différents d'un phonème. Les scores étudiés ci-dessous sont calculés en comptant 1 point par bonne réponse[56]. L'ensemble de l'étude est beaucoup plus raffinée que ne semble l'indiquer cette présentation rapide, puisqu'elle différencie les types de bonne réponse, ainsi que la position dans le logatome du phonème discriminant. Cependant le niveau d'analyse le plus grossier suffit pour ce que nous voulons mettre en évidence.

On étudie donc les scores 1N (maternelle, passation normale) et 1R (maternelle, passation rapide) selon les classes

Analyse de la variance de (score 1 N) selon les positions de (Classe)
sous le filtre (maternelle)

Classe	Effectif	Moyenne	Ecart-type
MS	46	14.54	3.99
GS	46	17.13	2.95
ENSEMBLE	92	15.84	3.74

F(1,90) = 12.21, s. à .01



Selon cette analyse classique, les élèves de grande section réussissent mieux que ceux de moyenne section pour 1N.

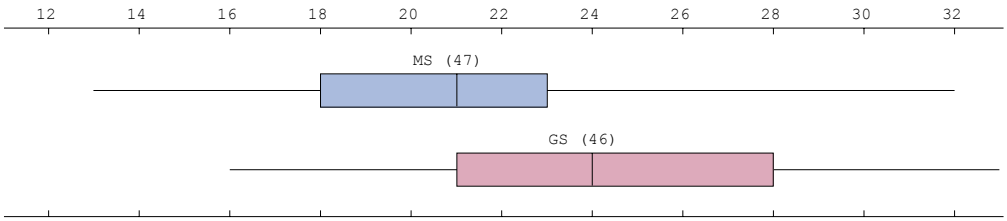
Analyse de la variance de (score 1 R) selon les positions de (Classe)
sous le filtre (maternelle)

Classe	Effectif	Moyenne	Ecart-type
MS	47	21.87	4.86
GS	46	24.46	4.36
ENSEMBLE	93	23.15	4.80

[55] Remerciements à ELODIE GUITTON ET CELINE MOREL, (2007) Mémoire présenté pour l'obtention du certificat de Capacité d'Orthophoniste. Institut d'Orthophonie Gabriel Decroix, Université de Lille II.

[56] « pas pareil » pour des logatomes différents, « pareil » pour des logatomes identiques. « Je ne sais » compte toujours pour une mauvaise réponse.

F(1,91) = 7.12, s. à .01



De même pour la passation 1R, où les scores sont d'ailleurs plus élevés de 7 points[57].

On pourrait s'en tenir là si on ne se souvenait à propos que la série comporte 36 paires de logatomes, et qu'un individu qui répondrait au hasard aurait en moyenne 18 bonnes réponses. On examine alors la distribution des scores des élèves de MS à la passation 1N.

(score 1 N)

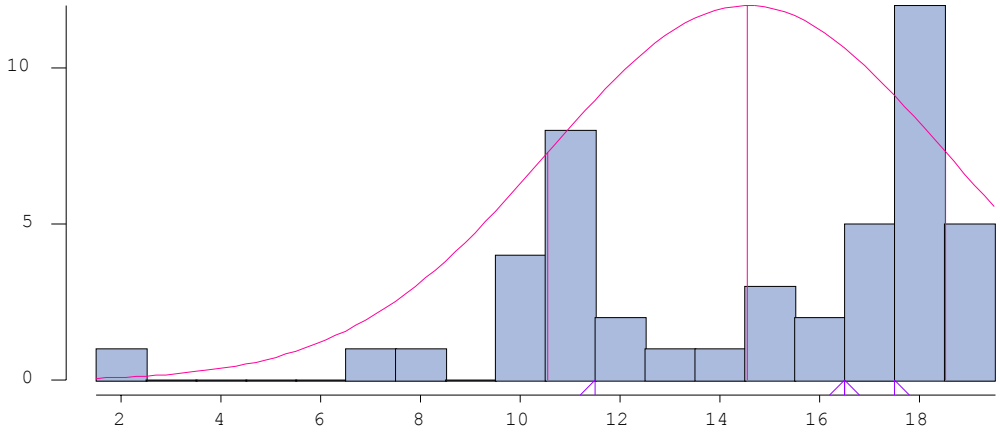
sous le filtre (MS)

	effectifs	%/Total	% cumulés
2	1	2.17%	2.17%
7	1	2.17%	4.35%
8	1	2.17%	6.52%
10	4	8.70%	15.22%
11	8	17.39%	32.61%
12	2	4.35%	36.96%
13	1	2.17%	39.13%
14	1	2.17%	41.30%
15	3	6.52%	47.83%
16	2	4.35%	52.17%
17	5	10.87%	63.04%
18	12	26.09%	89.13%
19	5	10.87%	100.00%
Total	46	100.00%	

Valeur modale : 18 (n=12)

Médiane entre 16 & 17

Moyenne 14.54, écart-type 3.99



On s'aperçoit que 5 élèves seulement ont obtenu mieux que le 18 qui serait le score dû au hasard. Et encore, le 19 lui-même est-il significatif ?

[57] On vérifie par ailleurs la significativité de cette relation en croisant les scores des deux passations et en considérant le test |t| de Student testant la différence de moyenne sur échantillons appareillés. Il est en l'occurrence très significatif.

Les lois des probabilités nous enseignent qu'un phénomène de fréquence théorique F, au cours de mesures répétées, se manifestera avec une fréquence moyenne f = F, et un écart-type s = F (1-F) / racine(N), où N est le nombre d'observations.

L'intervalle de confiance pour cette moyenne, au seuil P=.05, est donc [m-1.96s m+1.96s].

Cela signifie qu'au seuil choisi[58], les valeurs comprises dans l'intervalle de confiance ne différent pas significativement de la valeur moyenne. Seules les valeurs en dessous de m-1.96s peuvent être considérées comme significativement inférieures à la moyenne, seules les valeurs au dessus de m+1.96 s peuvent être considérées comme significativement supérieures à la moyenne.

Dans le cas étudié, la fréquence théorique et moyenne est de .5, qui correspond à des réponses au hasard sur deux issues (une chance sur deux).

L'écart-type dépend de l'effectif. Pour 46 sujets, il vaut $.5 \times (1-.5) / \sqrt{46}$, soit.037.

Les bornes de l'intervalle de confiance en fréquence sont donc .5-.037=.463 et .5+.037=.537

Rapportées à 36 essais, les bornes en nombre de réussites sont donc 16.77 et 19.33.

En d'autres termes, pour cet effectif total, les scores de 17 à 19 ne différent pas significativement du score au hasard.

Est-il raisonnable de considérer comme des réussites des scores qui ne différent pas significativement du score au hasard ? Certainement pas. Quant aux scores inférieurs à 17, ils ne peuvent être interprétés qu'en termes de réponses en partie systématiquement fausses.

Du coup, le tableau des scores 1N pour les MS se relit ainsi :

- 24 enfants, soit 52%, ont des résultats inférieurs à l'espérance aléatoire
- 22 enfants, soit 48%, ont des scores correspondant à des réponses au hasard
- aucun n'enfant ne répond mieux qu'au hasard

Qu'est-ce que ça signifie ? Au minimum que l'épreuve ne convient nullement aux enfants de moyenne section, soit qu'il n'en comprennent pas les consignes, soit qu'ils n'aient pas les capacités requises, soit que les conditions de passation les incitent à la confusion[59]. Rappelons aussi que les mauvaises réponses peuvent être des « je ne sais pas ».

La situation s'améliore à peine avec les élèves de grande section, puisque 3 enfants seulement obtiennent le score de 20, premier score significativement supérieur à celui du hasard.

Le tableau des scores 1N pour les GS s'établit alors ainsi :

- 11 enfants, soit 23%, ont des résultats inférieurs à l'espérance aléatoire
- 32 enfants, soit 70%, ont des scores correspondant à des réponses au hasard
- 3 enfants, soit 7%, répondent mieux qu'au hasard

Le bénéfice est surtout dans la diminution des résultats inférieurs à l'espérance aléatoire, qui est peut-être (ce serait à vérifier) liée à des « je ne sais pas » dus à l'inhibition.

On souhaite épargner ici au lecteur, autant que possible, une série fastidieuse de tableaux et d'histogrammes.

---

[58] La valeur de 1.96 correspond au seuil de .05 parce que dans la loi normale réduite 5% des effectifs sont situés à l'extérieur de la fourchette $\pm$ 1.96. Pour les seuils de probabilité de .10 et .01, on utiliserait respectivement les valeurs 1.65 et 2.58.

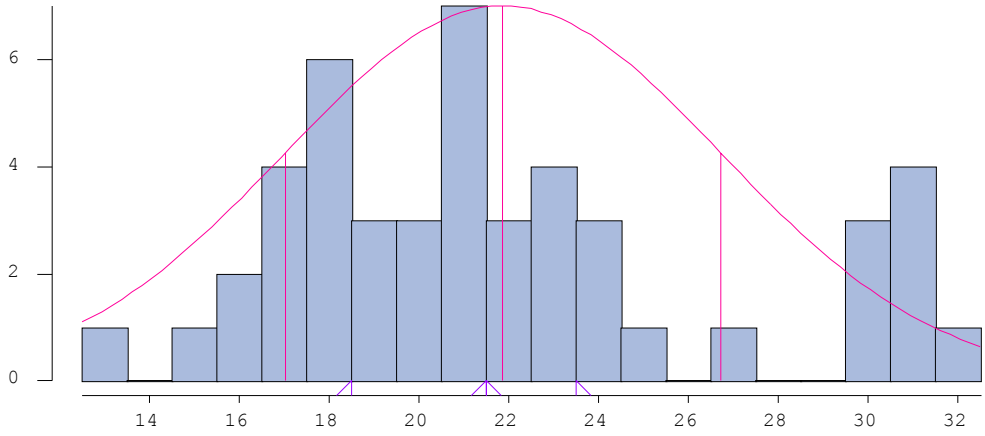
[59] On notera qu'alors que la réussite suppose, sauf coup de chance, que les trois conditions soient réunies : présence des capacités, compréhension des consignes, conditions de passation appropriées ; l'échec peut provenir du défaut d'une quelconque ou de plusieurs de ces conditions, sans qu'il soit possible, en tous cas sur la base des données disponibles, de déterminer la ou lesquelles.

Pourtant un coup d'œil sur la condition de passation Rapide n'est pas dépourvu d'intérêt :

(score 1 R)
sous le filtre (MS)

	effectifs	%/Total	% cumulés
13	1	2.13%	2.13%
15	1	2.13%	4.26%
16	2	4.26%	8.51%
17	4	8.51%	17.02%
18	6	12.77%	29.79%
19	3	6.38%	36.17%
20	3	6.38%	42.55%
21	7	14.89%	57.45%
22	3	6.38%	63.83%
23	4	8.51%	72.34%
24	3	6.38%	78.72%
25	1	2.13%	80.85%
27	1	2.13%	82.98%
30	3	6.38%	89.36%
31	4	8.51%	97.87%
32	1	2.13%	100.00%
Total	47	100.00%	

Valeur modale : 21 (n=7)
Médiane entre 21 & 22
Moyenne 21.87, écart-type 4.86



L'effectif est à peine différent (un enfant de plus), et les bornes entières de l'intervalle de confiance n'en sont pas affectées. Les grosses différences sont dans la répartition des scores :

- 4 enfants, soit 9%, ont des résultats inférieurs à l'espérance aléatoire
- 13 enfants, soit 17%, ont des scores correspondant à des réponses au hasard
- 30 enfants, soit 64%, répondent mieux qu'au hasard

Dans ce mode de passation, l'épreuve devient utile au regard de la population MS (à condition peut-être qu'il y ait eu auparavant une passation Normale, c'est à vérifier). L'effet est, comme on s'y attend, encore plus net chez les GS.

En résumé, toujours comparer ce qu'on obtient à ce que le hasard peut produire : c'est l'essence même de la statistique inductive.

Annexe 1 : Glossaire statistique

Ce glossaire est en partie issu de l'ouvrage de docimologie déjà cité, enrichi de quelques définitions préparées pour la notice d'EVALO 2.6.

analyse de variance:ou ANOVA. Démarche consistant à comparer les *distributions* partielles (les distributions pour les différents sous-ensembles de sujets) d'une variable numérique selon une variable catégorielle, pour déterminer si leurs moyennes sont significativement différentes. Utilise le *F de Snédécor-Fisher* pour rejeter ou non l'hypothèse nulle (= les moyennes en sont pas différentes). Peut être complété par l'analyse arborescente des *contrastes*.

arborescence des contrastes : Complément graphique de l'analyse de variance, qui présente sous forme de dendrogramme horizontal les coupures successives opérées sur une série de classes ordonnées par moyennes croissantes d'une mesure sur le groupe, en maximisant le *|t| de Student* et donc la différenciation des moyennes entre les deux membres séparés par la coupure. L'arborescence indique de manière récursive les dichotomies les plus significatives et porte à chaque bifurcation la mention du seuil de probabilité du |t|, qui permet de décider des rapprochements ou coupures significatifs entre classes.

bimodale : Se dit d'une courbe à deux *modes* (bosses, ou maxima locaux). Peut faire soupçonner le mélange de deux populations aux caractéristiques distinctes.

boîtes à moustaches : Traduction littérale de *boxes-and-whiskers*, représentation graphique associée à l'analyse de variance, avec pour chaque groupe une boîte centrale contenant la moitié de la population comprise entre le premier et le troisième quartile.

centrée-réduite : Transformation qui substitue à une valeur numérique la valeur obtenue en y soustrayant la *moyenne* de la distribution (centrage) avant de diviser par *l'écart-type* (réduction). Souvent désignée par la lettre **z**.

compétence : Disposition permanente et individuelle d'un individu, qui lui permet d'accomplir avec succès certaines tâches. Non observable directement, la compétence est induite de l'observation de *performances*.

corrélation : Tendance de deux grandeurs numériques à varier ensemble, de sorte que les valeurs élevées de l'une soient associées aux valeurs élevées de l'autre, et *vice versa*. Elle se mesure par le coefficient r de Bravais-Pearson (r_{BP}).

courbe en cloche : *Bell Curve* chez les anglophones, courbe représentative de la loi mathématique de Laplace-Gauss-Seidel, dite Loi Normale, associée aux événements issus du hasard.

croisement (variables) : Opération consister à trier une population statistique simultanément selon deux *variables*.

dépendance statistique : Situation d'une *variable* dont les valeurs peuvent être prédites plus ou moins précisément en connaissant les valeurs d'une autre variable. Antonyme : *indépendance*.

difficulté : La difficulté d'une *épreuve* ou d'un *item* est mesurée par le taux d'individus qui y échouent.

discrimination : Tendance d'un *item* à opposer nettement ceux qui, globalement, réussissent, et ceux qui, globalement, échouent.

dispersion : Tendance de la distribution d'une variable numérique à présenter des valeurs éloignées de la *moyenne*. La mesure de la dispersion est usuellement *l'écart-type*.

distribution : Ce terme désigne le détail du nombre de sujets correspondant à chaque valeur différente que peut prendre une variable : la distribution des âges dans une population majeure est le nombre de personnes qui ont 18 ans, 19 ans, 20 ans … jusqu'au maximum rencontré.

écart-type : Racine carrée de la *variance* d'une distribution numérique, utilisé comme indice de *dispersion*.

échelle d'intervalles : Un système de mesure tel que l'intervalle entre deux valeurs est comparables avec l'intervalle entre deux autres valeurs, autrement dit qu'il existe un système régulier d'unités.

épreuve : Situation *d'évaluation* standardisée en termes de durée, de lieu, d'exercices imposés, contribuant à un *examen,* à un *concours ou à un test*. Désigne également (mais on dit aussi le sujet) la description des exercices imposés dans ce contexte.

étalonnage : relevé des valeurs caractéristiques d'une mesure sur une population nombreuse servant de référence, en vue de pouvoir ultérieurement positionner rapidement un sujet par rapport à cette population. Si la distribution de la mesure est approximativement *normale*, l'étalonnage peut-être paramétrique, c'est-à-dire s'appuyer sur les paramètres que sont la moyenne et l'écart-type pour déterminer les seuils d'atypicité forte ou faible (moyenne plus ou moins un écart-type), voire les seuils de pathologie (moyenne moins deux écarts-types). Si, comme c'est très fréquent, la mesure n'est pas normale, la moyenne et l'écart-type ne sont pas des résumés utilisables des caractéristiques de la variable, et il faut passer à des méthodes d'étalonnage non-paramétriques, tels que la *normalisation*, le *quantilage*, ou encore le *seuillage* fragilité/difficulté.

F de Snédécor-Fisher : Statistique qui, dans une analyse de variance, correspond au rapport entre la variance interclasse (due à la répartition des sujets en classes) et la variance intraclasse (due au « bruit » des variations individuelles). Plus cette statistique est élevée, plus l'explication des différences entre sujets par leur appartenance aux classes est vraisemblable.

fiabilité : Ou cohérence : tendance des items composant un test à mesurer la même *compétence*, manifestée par des *corrélations* élevées entre *items*.

fidélité : qualité d'une mesure qui, lorsqu'elle est réitérée, est en forte corrélation avec elle-même. Tendance d'un test, en nouvelle passation, à fournir des résultats similaires à la première. Tendance d'un correcteur à évaluer à nouveau un travail avec une *notation* proche.La fidélité d'un test est parfois vérifiée au moyen d'une forme parallèle du test (d'autres items liés aux mêmes compétences), pour éviter les effets d'apprentissage (*test wiseness*), mais la vérification du parallélisme entre épreuves nécessite lui-même une expérimentation séparée.

histogramme : Représentation graphique associée au tri d'une variable numérique, dans laquelle les différentes valeurs sont rendues par une suite ordonnée de rectangles de hauteur proportionnelle aux effectifs associés.

indépendance : Situation de deux variables entre lesquelles aucune relation ne peut être démontrée.

item : Le plus petit élément d'une *épreuve*, pouvant faire l'objet d'une *évaluation* individuelle.

khi2 (χ^2): Coefficient de contingence associé au croisement de deux variables nominales. La significativité du khi2 s'apprécie en fonction du nombre de degré de liberté du tableau croisé (l-1)(c-1), l et c étant les nombres de lignes et de colonnes.

loi normale : Loi statistique décrivant les phénomènes résultant de l'addition d'un grand nombre de petits phénomènes indépendants et de faible amplitude. Caractérisée par une courbe en cloche, symétrique, ventrue au centre et effilée aux extrémités, elle est le modèle d'une distribution aléatoire, représentation mathématique du hasard.

médiane : Coupure au sein d'une distribution numérique ou ordinale, telle qu'environ la moitié de la population se trouve de part et d'autre.

mode: Dans une distribution numérique, valeur pour laquelle l'effectif est le plus grand ; correspond au pic graphique dans *l'histogramme* associé.

moyenne : Indicateur de tendance centrale des distributions numériques, calculé comme la somme de toutes les valeurs, divisée par le nombre de valeurs.

multimodale : Se dit d'une distribution numérique comportant plusieurs modes, et associée à un *histogramme* à plusieurs maxima locaux.

normalisation : l'une des transformations non-paramétriques d'une distribution aux fins d'étalonnage. Comme les autres transformations non-paramétriques (*quantilage*, *seuillage*), elle nie la qualité métrique de la variable employée pour ne conserver que sa qualité ordinale : Les cases de l'étalonnage, en nombre fixé par convention, sont remplies dans l'ordre des valeurs de la variable d'origine, jusqu'à atteindre les effectifs souhaités : aucune formule mathématique ne résume donc la transformation, qui est contingente aux caractéristiques de la population de référence, et qui engendre une *table de conversion*. Dans le cas de la normalisation (souvent en un nombre impair de cases, de 5 à 11 usuellement), les effectifs attendus sont ceux qu'on obtiendrait sous une Loi Normale (ventru au milieu, effilé aux extrémités). La forme symétrique des distributions de QI est due à une normalisation et nullement à une caractéristique propre de la mesure sous-jacente. On reproche parfois à la normalisation de contribuer à accréditer la forme « normale » comme forme à laquelle on doit s'attendre dans toutes les situations de mesure, ce qui est tout à fait erroné, puisque la loi Normale correspond aux mesures résultant de l'accumulation d'un grand nombre de petits phénomènes indépendants (passage à la limite d'une loi binomiale symétrique), ce qui n'a aucune raison particulière d'être le cas des tests et épreuves.

normalité : Une distribution est normale si sa distribution de fréquences est conforme à la loi de Laplace-Gauss, dite Loi Normale, parfaitement symétrique, ventrue au milieu, effilée aux extrémités, dont la jolie formule figure dans tous les manuels de statistiques. La normalité d'une ou plusieurs distributions est souvent une condition nécessaire à la légitimité de l'utilisation de certains instruments statistiques, notamment les approches paramétriques utilisant la moyenne et l'écart-type, mais aussi la corrélation entre deux variables.

nuage de densité : Représentation graphique associée au croisement de deux variables numériques. Des taches plus ou moins foncées correspondent à des concentrations de population dans le plan défini par les variables.

performance : Actes et travaux effectivement réalisés par un individu dans une circonstance donnée, et dont on essaie d'induire sa *compétence* dans le domaine correspondant.

pertinence (d'un item) : Relation sémantique valable entre un item et ce qu'il est censé mesurer : degré d'appropriation à son objet.

prédictivité : capacité d'un test passé à un temps t_0 à prédire les résultats d'un autre test ou de toute autre mesure prise sur les mêmes individus à un temps t_1 ultérieur. Elle s'apprécie en même temps que la sensibilité et la spécificité, sur un tableau 2×2 croisant la prédiction (non ou oui : p0, p1) et la vérification (non ou oui : q0, q1). Si on désigne par p0q1, par exemple, le nombre de cas où la chose à prédire, quelle qu'elle soit, n'a pas été prédite (p0), mais s'est réalisée (q1), les formules sont les suivantes : la sensibilité se mesure par p1q1/(p1q1+p0q1), la spécificité par p0q0/(p0q0+p1q0), la prédictivité par (p0q0+p1q1)/(p0q0+p0q1+p1q0+p1q1). Ces statistiques ne peuvent faire l'objet d'une interprétation valide que si le test du khi² sur le tableau croisé est significatif. La sensibilité s'apprécie alors comme la capacité du test à réellement repérer les sujets pour lesquels la chose se réalisera, la spécificité comme la capacité du test à écarter ceux pour lesquels la chose ne se réalisera pas, la prédictivité générale comme la capacité à faire des prédictions justes par rapport à l'ensemble des prédictions.

quantilage : une transformation non-paramétrique, concurrente de la normalisation, à laquelle elle ressemble beaucoup, à ceci près que les effectifs souhaités sont équiprobables : on recherche l'ajustement à une distribution plate. Le nombre de cases

décline le type de quantilage : cent cases font des centiles, dix des déciles, cinq des quintiles, quatre des quartiles, trois des tertiles. Avec deux cases, on rejoint la dichotomie sur la médiane. On rencontre aussi des quantilages à 11 cases (0 à 10) ou 21 cases (0 à 21), qui créent un risque de confusion avec des notes scolaires traditionnelles.

rang : Ordre de classement des individus rangés par valeur de *performance* décroissante.

relation d'ordre : Dans un système de valeurs, la relation d'ordre est ce qui permet de dire que tel élément est avant ou après tel autre.

reproductibilité : Au sens de Guttman, tendance d'une série *d'items* à pouvoir être ordonnés selon un schéma pyramidal, le succès à un item impliquant le succès à tous ceux qui le précèdent dans la série.

rhô (ρ) de Spearman : coefficient de *corrélation* alternatif à celui de Bravais-Pearson, quand les qualités métriques des variables étudiées sont douteuses, notamment si elles ne sont pas approximativement *normales*. Le coefficient de Spearman se calcule sur les rangs des sujets dans la population et non sur les valeurs brutes.

seuil (de probabilité) : Taux de risque d'erreur que l'on accepte en prenant une décision statistique, et qui dépend de la valeur des tests statistiques. Les seuils utilisés en Sciences Humaines sont le plus souvent .10 , .05 , .01, correspondant respectivement à des taux d'erreur de 10, 5 et 1%. Sur de grands effectifs, on utilise aussi les seuils .001, .0001 et .0000, qui correspondent à des taux d'erreur de 0,1%, 0,01%, et, pour le dernier, à un taux d'erreur quasi-nul (pas de décimale significative avant la cinquième).

seuillage : une transformation non-paramétrique concurrente de la normalisation et du quantilage, et dans laquelle on se préoccupe seulement de repérer en dessous de quel seuil d'une mesure on trouve 9% d'une population, puis 25%. Les 9% les plus faibles sont dits « en difficulté », les 16% suivants pour atteindre 25% en tout sont dits « fragiles ». D'autres seuils peuvent être employés, mais ceux-là sont bien adaptés à une démarche de dépistage en première intention, par exemple pour orienter vers des soins ou vers une remédiation plus légère.

|t| de Student : statistique qui évalue la différence entre une moyenne et une norme, ou entre deux moyennes, en fonction des effectifs et des écarts-types. Comme comparaison à une norme, il peut servir à vérifier si un échantillon s'écarte significativement des valeurs (normes) mesurées sur une population parente, et donc s'il est représentatif. Entre deux échantillons indépendants (deux groupes de sujets distincts sur une même mesure), il sert à déterminer si les deux moyennes relevées sont significativement différentes : il joue alors sur deux classes le même rôle que le *F de Snédécor-Fisher* sur un nombre quelconque de classes. D'ailleurs, dans ce cas, $F = t^2$. Entre deux échantillons appareillés (mêmes sujets, deux mesures), il peut servir par exemple à mesurer des effets de progrès ou d'apprentissage.

table de conversion : Double liste de valeurs permettant de passer des valeurs brutes d'une épreuve aux indices (numéros des cases) des *étalonnages* par *normalisation*, *quantilage* ou *seuillage*. Cette table est indispensable, car les transformations non-paramétriques ont en commun d'être arbitrairement liées au comptage des cas dans une population de référence et non à une formule mathématique de transformation.

test : Ensemble d'épreuves destiné à mesurer une capacité ou plusieurs. Terme utilisé surtout en psychologie ; les *tests*, élaborés et vérifiés scientifiquement, s'opposent aux simples exercices scolaires.

test-wise : Se dit d'un individu habitué à passer des tests et qui développe ainsi un *méta-savoir* propre à lui permettre de meilleures performances.

tri(d'une variable) : Ventilation d'une population selon les différentes valeurs possibles d'une variable : comptage des effectifs associés à chaque valeur.

validité : La validité d'un test ou d'une épreuve repose sur sa cohérence interne, sur sa vérification par comparaison à d'autres instruments et sur sa pertinence au regard de l'objet de la mesure.

variable : Toute entité capable de prendre une et une seule valeur pour un individu donné, et des valeurs éventuellement différentes pour des sujets différents.

variable ordinale : *Variable* dotée d'une relation d'ordre, comme le système de mention à un examen, souvent par opposition à une variable nominale, telle que la nationalité, pour laquelle l'ordre de présentation est nécessairement arbitraire (penser aux jeux olympiques où l'ordre des délégations nationales dans le défilé initial dépend des traductions des noms de pays dans la langue locale et son alphabet).

variance : Pour une distribution numérique, c'est la *moyenne* des carrés des écarts à la moyenne. On en déduit *l'écart-type* en extrayant la racine carrée.

Annexe 2 : Brève présentation d'Hector

Hector, dans sa plus récente version Hector² (depuis le printemps 2008), est l'avatar présent d'une série[60] de logiciels d'analyse de données que l'auteur a développés et qu'il a utilisés pour la recherche et l'enseignement.

D'un point de vue technique, Hector se compose d'un *noyau* riche et amplement suffisant pour les travaux statistiques en sciences humaines jusqu'aux frontières du doctorat, et *d'extensions* plus sophistiquées, développées selon les besoins et demandes des chercheurs confirmés. L'auteur est un ancien enseignant, un toujours chercheur, et, nécessairement aussi un ingénieur en logiciels ; mais ce n'est pas un commerçant. Aussi Hector fait-il l'objet d'un modèle économique spécifique :

- la version de base d'Hector est gratuite : elle se télécharge à l'adresse http://alain.dubus.r.et.d.free.fr , ainsi que sa documentation et des corpus d'exemples. Les étudiants et les professionnels (à des fins d'autoformation) sont invités à visiter régulièrement ce site, car Hector est un logiciel vivant, dont les mises à jour sont fréquentes. L'usage professionnel de cette version gratuite est considéré comme déloyal, quoique l'auteur ne dispose d'aucun moyen de rétorsion et n'entend pas consacrer fût-ce une minute de son temps à faire la police. De plus, les utilisateurs de cette version ne peuvent prétendre à aucun soutien technique ou méthodologique direct de la part de l'auteur.
- la version professionnelle-recherche n'est pas non plus à vendre, mais elle peut être mise à disposition de laboratoires, d'équipes de recherche, d'UFR ou d'Instituts dans le cadre de conventions d'assistance méthodologique à la recherche ou à la formation. Ces conventions sont assorties d'une rémunération de l'auteur à des niveaux qui paraîtront bien modestes à quiconque a fait l'acquisition d'un des coûteux ensembles d'analyse de données disponibles sur le marché. Pour tous renseignements, s'adresser à adubus.rechdev@laposte.net L'assistance méthodologique s'exerce sous forme d'une *hot-line* en courrier électronique.
- de temps à autre, dans le cadre d'une campagne de recherche importante, ce qu'on appelle « un grand compte », c'est-à-dire un organisme de recherche doté de moyens substantiels, demande et finance un développement spécifique de nouvelles fonctionnalités, qui sont par la suite intégrées à la version professionnelle.

La présentation faite ici d'Hector ne vise pas à remplacer les manuels du logiciel, qui sont disponibles au format PDF sur le site, mais à donner une idée de ce que peut faire le logiciel et à permettre au lecteur de se représenter comment le logiciel peut satisfaire ses besoins.

[60] Depuis 1978-79, pour les besoins de sa propre thèse, quelques logiciels anonymes, puis, au fil des années pour les besoins de la recherche et de l'enseignement, et en tenant compte des évolutions techniques des ordinateurs individuels, TEL3, Occam, Guillaume, Adso de 1.0 à 3.3, Nestor, Hector, et même un Viktor resté à l'état de chantier.

La version de base d'Hector

Principes

Le modèle sous-jacent à Hector est celui du *corpus* de données, qu'on peut se représenter comme une grille rectangulaire où les lignes représentent des *variables* et les colonnes des *sujets* : c'est en tous cas comme cela qu'est organisée la grille de saisie des données. Ce modèle est très proche de celui des tableurs, à ceci près que la tendance des tableurs est plutôt à placer les variables en colonnes et les sujets en lignes, et de celui des bases de données, où les variables sont plutôt appelées *champs*.

Du point de vue des bases de données, un corpus Hector est une seule table, et Hector ne conserve pas trace permanente de relations entre différentes tables au sein d'une même base de données, mais un mécanisme simple et robuste du langage des formules permet dans Hector de faire fonctionner de véritables bases de données relationnelles multi-tables.

Avec les systèmes de base de données, Hector partage un fort souci de typage rigoureux des variables. En effet, du type d'une variable dépendent les opérations statistiques légitimes qu'on peut lui appliquer. Les types de données dans Hector sont les suivants :

- Le type *numérique* : il code des nombres entiers ou non, positifs ou négatifs
- Le type *calendaire* : variante du précédent, il code des dates.
- Le type *logique* : il code en Vrai ou Faux
- Le type *ordinal* : il code des catégories, ou classes, dont les valeurs sont représentés par de courts textes, les *étiquettes*. L'ordre des catégories a une signification.
- Le type *nominal* : il code aussi des catégories par des étiquettes, mais sans ordre significatif entre elles.
- Le type *texte libre* : il peut contenir des textes de longueur quelconque et n'autorise aucun traitement statistique direct, mais peut être manipulé par le langage des formules pour en extraire des variables utilisables.

Les étapes préalables au traitement des données

Deux grands cas de figure peuvent se présenter :

- Les données sont disponibles sous forme papier, dans des questionnaires ou des protocoles de test : elles n'ont pas encore été saisies de manière informatique
- Les données ont déjà été saisies, généralement dans un tableur ou un traitement de texte

Le second cas est le moins favorable, car on n'a pas bénéficié des outils de saisie d'Hector et des moyens de contrôle qu'ils fournissent. On devra dans ce cas se débrouiller pour extraire les données dans un format rustique mais universel, dit *texte tabulé*, qu'on pourra ensuite importer dans Hector pour constituer un corpus.

Le premier cas est le cas normal, où l'on commence par décrire les propriétés des variables avant d'en saisir le contenu. Les étapes sont alors les suivantes :

Création du corpus et plan de codage

Pour chaque variable, on décide de son type et de son intitulé. Pour les numériques, on décide également des valeurs minimales et maximales qu'elles peuvent prendre, ainsi que de leur précision (le nombre de chiffres après la virgule). Pour les ordinales et les nominales, on décide aussi de la liste des étiquettes représentant les valeurs admissibles.

Ces décisions sont de grande importance, même s'il est possible de revenir dessus par la suite. Il est très important de choisir des intitulés de variables intelligibles, non pas pour Hector qui n'en a cure, mais bien pour l'usager lui-même. Il en va de même des

étiquettes. Il est surtout très important de ne pas se censurer dans le codage, et de prendre en compte toute l'information dont on dispose.

Par exemple, il est déconseillé de créer des classes d'âge *a priori*, alors qu'il est extrêmement facile de les fabriquer automatiquement par la suite à partir d'âges détaillés, tandis que la finesse de l'information qui n'est pas prise en compte à cause d'une catégorisation prématurée est perdue à jamais.

De plus, les limitations imposées aux variables dans la description du plan de codage vont permettre des mécanismes de contrôle des erreurs dans la saisie.

La saisie manuelle des données

Elle se fait dans une grille où les variables sont en lignes et les sujets en colonnes. Si la saisie commence en cliquant avec la souris dans la case où l'on désire commencer, elle se peut se poursuivre en employant uniquement les touches du clavier, avec une ergonomie conservée même dans le cas d'un clavier d'ordinateur portable.

Chaque case, où l'on saisit révèle une zone d'édition spécialisée selon le type de la variable : si les dates sont tapées au format jj/mm/aaaa, et les nombres frappés en clair, les logiques sont une simple bascule entre 'V' pour Vrai et '-' pour Faux (contraste visuel maximal), et les variables à étiquettes affichent la liste des valeurs possibles, liste qu'on parcourt avec les touches fléchées du clavier avant de valider avec Entrée.

Toute tentative pour entrer une valeur erronée[61] provoque un message d'erreur. Si on s'aperçoit qu'on a fixé des limites trop étroites pour une variable numérique, il est toujours possible d'aller modifier ces limites avant de reprendre la saisie.

La saisie peut évidemment être effectuée en plusieurs fois. De même, la notion de corpus « achevé » n'existe pas sous Hector, car il est toujours possible d'ajouter de nouveaux sujets et de nouvelles variables.

La saisie manuelle peut être complétée par une importation de fichier texte, et *vice versa*. Des protocoles de saisie répartie [62]sont décrits dans les manuels.

Le traitement des données dans la version de base, hors collections

On expose ici rapidement, du plus simple au plus complexe, les différentes strates de traitement proposées.

Traitement d'une variable à la fois

Le traitement d'une seule variable s'appelle aussi le *tri à plat*. Il peut prendre des formes différentes selon le type de la variable[63], mais offre toujours quatre éléments dont on peut empêcher ou autoriser séparément l'affichage :

- Le résultat graphique ou *graphe*
- Le résultat tabulaire, ou *table*
- La statistique *locale* (ou de détail)
- La statistique *globale*

Pour chaque type de variable, Hector propose une présentation par défaut, mais un panneau d'options est accessible, où d'autres choix peuvent être faits. Ces choix seront conservés pour une autre session de travail dans un fichier d'habitudes.

Ainsi, pour une variable numérique,

[61] Date impossible, comme un 30 Février ou un nombre hors des limites fixées au préalable.

[62] Plusieurs personnes saisissent des parties du corpus séparément avant assemblage. Cette fonctionnalité, très utiles pour les binômes d'étudiants en orthophonie, devient indispensable pour les recherches plus volumineuses et réparties sur plusieurs territoires.

[63] calendaire / numérique / logique / ordinale / nominale / libre

- Le graphe est un histogramme, sur lequel sont projetés optionnellement le tracé de la distribution normale de mêmes paramètres et les emplacements des quartiles.
- La table est la double liste des valeurs possibles et des effectifs correspondants. Si le nombre de valeurs possibles est trop élevé[64], les valeurs sont regroupées en classes.
- La statistique locale complète la table par une colonne des pourcentages et une colonne des pourcentages cumulés.
- La statistique globale est composée du mode , de la médiane, de la moyenne et de l'écart-type, auxquels s'ajoutent optionnellement le point de coupure contrastée[65] et les statistiques de normalité[66] et de symétrie.

Autre exemple, pour une variable nominale,

- Le graphe est par défaut un diagramme en secteur[67], mais optionnellement un graphe en barre ou en couches.
- La table est la double liste des étiquettes des valeurs présentes et des effectifs correspondants.
- La statistique locale complète la table par une colonne des pourcentages.
- La statistique globale propose le coefficient d'efficacité entropique[68]

Pour tous les graphes du tri à plat, comme d'ailleurs pour ceux des croisements, les options offrent le choix entre couleurs vives, couleurs pastels, dégradés de gris, hachures ou douze gammes de camaïeux[69].

Traitement de deux variables à la fois

Cela s'appelle aussi le *croisement*. Là aussi on retrouve le graphe, la table et les statistiques locales et globales selon le type des variables, mais la combinatoire des types génère une plus grande variété de situation.

On peut distinguer les cas « homogènes », où les deux variables sont de même type, et les cas « hétérogènes », où les deux variables sont de types différents.

Pour le croisement homogène de numériques (ou de calendaires),

- Le graphe est un nuage de densité, qu'un certain nombre de paramétrages optionnels permettent d'ajuster au mieux aux caractéristiques des données traitées.
- La table est le tableau croisé des effectifs selon les valeurs des deux variables.
- La statistique locale n'apporte ici rien de particulier
- La statistique globale est, par défaut, le coefficient de corrélation r de Bravais-Pearson, et, optionnellement, le coefficient de corrélation par rangs rhô de

[64] Au regard de limites fixées dans le panneau des options

[65] Le point où il faudrait trancher dans la distribution pour obtenir deux sous-ensembles les plus contrastés possibles selon le |t| de Student.

[66] La normalité d'une distribution numérique est un élément important car c'est une condition nécessaire à la légitimité de certaines autres opérations statistiques.

[67] Encore appelé camembert.

[68] Mesure de l'homogénéité de la distribution, c'est-à-dire de la tendance des différents effectifs à être égaux entre eux.

[69] Ça ne sert effectivement pas à grand chose, mais un jour l'auteur, fatigué des statistiques d'aspect austère, s'est fait plaisir en jouant avec les couleurs.

Spearman[70], la covariance et les coefficients de régression linéaire, et enfin le |t| de Student sur échantillons appareillés[71].

Pour le croisement homogène de logiques,

- Le graphe est le schéma en colonnes et pourcentages
- La table est le tableau de contingence des nominales (cf. *infra*)
- La statistique locale est celle du tableau de contingence
- La statistique globale offre le jeu le plus fourni[72] de tous les cas de figure, avec optionnellement les coefficients de corrélation de Bravais-Pearson et de Spearman, le Khi2 et le coefficient normé de contingence, le test de l'implication, le coefficient gamma de coordonnancement de Goodman-Kruskal et les paramètres de sensibilité, spécificité et prédictivité.

Pour le croisement homogène d'ordinales,

- Le graphe est le schéma en colonnes et pourcentages
- La table est le tableau de contingence
- La statistique locale est celle du tableau de contingence
- La statistique globale est celle des croisements de nominales (cf. *infra*), enrichi du coefficient de corrélation de Spearman et du gamma de Goodman-Kruskal.

Pour le croisement homogène de nominales

- Le graphe est optionnellement la représentation de l'Analyse Factorielle des Correspondances Simple[73], ou le schéma en colonnes et pourcentages.
- La table est le tableau de contingence, c'est-à-dire le tableau croisé des effectifs selon les étiquettes des deux variables
- La statistique locale comporte optionnellement les pourcentages-ligne et les pourcentages-colonne, ainsi que le signe des associations locales[74] et la couleur des associations locales[75].

De manière générale, les croisements hétérogènes se rallient aux modalités de croisement homogène de moindre *qualité* de mesure, celle-ci décroissant des numériques-calendaires

---

[70] Dont l'usage est préféré à celui de Bravais-Pearson quand l'une au moins des deux variables n'est pas normale.

[71] Particulièrement utile quand il s'agit de comparer des scores avant / après, à l'occasion d'une séquence d'apprentissage.

[72] Cette richesse particulière ne doit pas étonner, la variable logique étant en quelque sorte la variable « parfaite », qui réunit aux siennes propres (Vrai/Faux et toutes les opérations logiques) les qualités des numériques (0/1) et des catégorielles (Oui/Non), ce qui autorise une grande variété d'opérations. De plus, la variable logique est à la base du fonctionnement même des ordinateurs, et fournit toujours les opérations les plus efficaces. Pour ces raisons, il est souvent avantageux de coder sous cette forme les informations élémentaires dont on dispose, sachant qu'il est toujours possible de les recombiner ensuite sous d'autres formes plus synthétiques.

[73] Cette technique souvent spectaculaire est principalement utilisée comme instrument d'exploration des données, et on la retrouve à ce titre, simple et multiple, dans les fonctionnalités de la version professionnelle, mais elle est ici proposée comme illustration du tableau de contingence. Son interprétation réclame toutefois quelques compétences, et, selon le public auquel on s'adresse, on pourra être amené à lui préférer l'autre représentation, plus simple de lecture.

[74] Une statistique qui repère les cases du tableau dont les effectifs sont les plus éloignés dans un sens ou dans l'autre des valeurs attendues sous l'hypothèse d'indépendance des variables, et qui signale ces déformations par un jeu de un, deux ou trois signes + ou -, selon que la déformation est significative à .10, .°05 ou .01.

[75] Les associations locales sont renforcées par des couleurs de fond de case elles-mêmes paramétrables optionnellement.

vers les nominales, en passant par les logiques et les ordinales. Ainsi un croisement entre une logique et une nominale sera-t-il traité comme un croisement homogène de nominales.

Cette règle connaît une exception majeure : le croisement mixte, ou ANOVA[76]. Celui-ci intervient quand l'une des variables est numérique ou calendaire, et l'autre catégorielle (logique, ordinale ou nominale).

La signification du croisement change légèrement dans cette situation, où il s'agit de tester l'hypothèse selon laquelle les moyennes observées pour la variable numérique selon les différentes valeurs de la variable catégorielle doivent être considérées comme significativement différentes ou au contraire être considérées comme de simples variantes d'une même valeur[77].

Dans ce cas de l'ANOVA,

- Le graphe est la boîte à moustaches[78], optionnellement mono- ou polychrome
- La table affiche, pour chaque catégorie, l'effectif, la moyenne et l'écart-type
- La statistique locale est ici l'arborescence des contrastes[79], dendrogramme reflétant les dichotomies successives maximisant le |t| de Student et affichant la significativité de celui-ci
- La statistique globale est le F de Snédécor-Fisher, rapport entre la variance inter-classes et la variance intra-classes. Optionnellement, le détail du calcul du F est affiché.

Traitement de trois ou quatre variables à la fois

La page des traitements simples autorise jusqu'à quatre variables à croiser simultanément. Toutefois, toutes les combinaisons ne sont pas autorisées, et notamment le système ne peut contenir plus de deux variables numériques-calendaires[80].

De manière générale, le traitement est celui des croisements de deux dernières variables, étudiés séparément sous condition de chacune des valeurs de la première ou de chaque combinaison de valeurs des deux premières. Ce traitement n'est réellement intéressant qu'à condition de disposer de gros effectifs, car l'accumulation des conditions tend à fractionner l'effectif en groupes trop petits pour être statistiquement utiles.

Une exception à cette organisation intervient à nouveau quand une seule des variables du système est numérique-calendaire, confrontée à deux variables catégorielles : on se retrouve alors dans la situation de MANOVA[81], où sont étudiées non seulement l'influence de chacune variable catégorielle, mais également celle de leur interaction. Le graphe spécifique à cette situation est plus facile à comprendre qu'à décrire abstraitement ici. Aussi s'en abstiendra-t-on. Si une quatrième variable catégorielle est introduite, la MANOVA est effectuée sous condition successivement de chacune de ses valeurs.

Traitement et filtres

Les filtres, disponibles partout dans Hector, mais qu'on présente ici pour la première fois, offrent un moyen très puissant d'organiser des traitements fins.

---

[76] Analysis Of VAriance, analyse de variance en français.

[77] Cette démarche est constante en docimologie comme en analyse des tests, puisqu'il s'agit de comparer différents groupes et de trancher s'ils obtiennent ou non des scores similaires.

[78] Traduction littérale de *box-and-whiskers*, système de boîtes horizontales prolongées par des traits et portant trace, pour chaque groupe de l'étendue de la distribution et des positions des quartiles.

[79] A vrai dire, il ne s'agit pas réellement d'une statistique locale, au sens où elle peut être interprétée indépendamment de la statistique globale : la variable étudiée n'est plus la catégorielle d'origine, mais des regroupements optimisés de ses valeurs, qui construisent virtuellement d'autres variables. Les |t| peuvent donc être interprétés même si le F global n'est pas significatif.

[80] Parce que l'auteur n'a pas su trouver ou imaginer une manière intelligente de représenter de telles combinaisons.

[81] Ou analyse de variance à plusieurs facteurs.

Un filtre est une variable logique[82] qui, une fois installée comme filtre et aussi longtemps qu'elle y est, restreint pour tous traitements la population à l'ensemble des sujets qui ont pour cette variable logique la valeur Vrai. Le reste de la population n'est bien sûr pas détruit, mais momentanément occulté.

Tous les résultats calculés sous un filtre sont affichés avec l'avertissement que ce filtre était en action, de manière à prévenir toute confusion.

Traitements élémentaires sur des collections de variables

Les collections de variables sont une manière d'organiser les variables en ensembles de même type, pour pouvoir effectuer sur ces ensembles des traitements spécifiques ou pour accélérer et démultiplier des traitements classiques.

Les collections se montent et se démontent très aisément dans la page de gestion des variables : il suffit de sélectionner en même temps plusieurs variables du même type, d'actionner le bouton approprié et de donner à la nouvelle collection un intitulé original, de préférence lisible et significatif.

Les tris en série

Une des premières activités du traitement des données, notamment dans la phase de vérification de vraisemblance[83], est de trier à plat toutes les variables. Ce peut-être extrêmement fastidieux et consommer pas mal de papier, et les tris en série proposent pour des collections de variables des formats de présentation beaucoup plus synthétiques et dépourvus de graphes comme de statistiques.

Diverses options permettent d'afficher des pourcentages cumulés, d'afficher un trait de séparation correspondant à la médiane quand le type des variable s' y prête, de trier les variables dans l'ordre des médianes croissantes, de fixer le nombre de colonnes affichables pour des numériques, de manière à obtenir des classes de valeurs pour les variables de grande étendue.

L'analyse des tests

Cet ensemble de fonctionnalités, largement utilisé dans le présent guide, rassemble les outils de base de la validation interne des tests :

- Discrimination, utilisable uniquement avec des variables logiques ou des numériques restreintes[84] aux valeurs 0 et 1, raison pour laquelle il est si utile de préférer ce type de codage quand on a le choix[85]. Optionnellement, le modèle de Gutmann confronte les données à un modèle hiérarchique d'implication des réussites.
- Cohérence et fiabilité, par le calcul des corrélations item-test et de l'alpha de Cronbach[86].

---

[82] Qui peur être une variable d'origine saisie initialement, ou, le plus souvent, une variable calculée au moyen du langage des formules pour décrire un sous-ensemble intéressant de la population, comme par exemple les enfants de 2 ans 6 mois à 3 ans 6 mois.

[83] Le moment où l'on s'assure qu'on n'a pas saisi des valeurs inattendues ou aberrantes.

[84] Et donc déclarées, dans le plan de codage, avec un minimum de 0, un maximum de 1 et une précision de 0 décimales.

[85] Assez typiquement, on rencontre des codages du type « 2 points pour réussite spontanée, 1 point pour réussite avec aide, 0 points pour échec ». Il est préférable de disjoindre en deux variables logiques « finalement réussi » et « a reçu une aide », à partir desquelles il est toujours possible de reconstituer précisément le détail qu'on souhaite, et qui ont le mérite d'autoriser les tests de discrimination.

[86] Ces statistiques ont été assez largement illustrées dans le corps du Guide, et on s'épargne donc de les décrire de nouveau.

Les matrices de statistiques

Cette fonctionnalité calcule *tous* les croisements deux à deux de deux collections de variables[87] : si l'une des collections contient 10 variables et l'autre 8, cette fonction calcule 80 croisements, et affiche sous forme de tableau (la *matrice*) la statistique globale appropriée au type des variables, selon les mêmes principes que dans les croisements simples : r de Bravais-Pearson, rhô de Spearman et |t| sur échantillons appareillés pour les numériques, khi2 en coefficient normé de contingence pour les nominales … F pour les croisements mixtes, etc.

Ces tableaux parfois très grands sont très riches en information, parfois trop riches. Deux types de dispositifs optionnels visent à en faciliter la lecture :

- L'un repose sur un système d'astérisques : trois pour une statistique significative à .01 ou moins, deux pour .05, une pour .10, aucune si non-significatifs.
- L'autre consiste à ne rendre visible que les statistiques significatives à un seuil à choisir parmi les trois précédents, auxquels s'ajoutent les seuils plus fins[88] de .001, .0001 et .0000.

Quand la matrice de statistiques concerne une collection croisée avec elle-même et d'un type tel que l'un au moins des coefficients de corrélation (r ou rhô) soit légitime, une option propose l'arbre des parentés, déjà présenté dans ce guide comme instrument d'étude de la structure interne des tests.

La création de nouvelles variables avec le langage des formules

Les données qu'on a saisies ou importées ne se présentent pas toujours sous le format le plus commode pour les opérations qu'on veut mener, et parfois on a besoin de combiner les variables entre elles pour en construire de nouvelles.

Cela se fait au moyen du langage des formules[89], analogue dans son principe aux systèmes employés dans les tableurs pour définir dans une case une valeur calculée sur la base des valeurs d'autres cases, mais en beaucoup moins compliqué[90].

Ainsi une variable peut-elle être calculée comme la différence de deux autres variables :

```
# durée_du_contrat
: date_de_fin - date_de_début ;
```

cette formule crée une nouvelle variable numérique[91] intitulée « durée du contrat », calculée d'un seul coup pour tous les sujets comme la différence entre les deux variables présumées calendaires « date de début » et « date de fin ».

Autre exemple, à condition d'avoir au préalable créé une collection « Notes en mathématiques » regroupant les variables qui représentent les notes obtenues aux divers devoirs du trimestre, la moyenne se calcule ainsi[92] :

---

[87] Eventuellement d'une collection avec elle-même.

[88] Cette disposition est rendue indispensable par l'analyse de gros corpus de données, tels que les évaluations des formations IUFM, qui possèdent des effectifs si importants que la moindre relation est extrêmement significative, ce qui impose d'être nettement plus exigeant en matière de significativité. Le seuil de .0000, qui signifie une probabilité si faible qu'aucune décimale n'apparaît avant la cinquième, est considéré comme l'équivalent pratique de la certitude, définie comme un risque nul de se tromper.

[89] Il a remplacé à partir du printemps 2008 le langage de dérivation antérieur, qui était trop complexe.

[90] Les langages des formules des tableurs se caractérisent notamment par une débauche de parenthèses et de point-virgules qui rendent très vite illisible la moindre formule. On note aussi la différence fondamentale entre l'usage d'adresses (cases ou colonnes) dans les tableurs pour repérer des entités, et l'usage d'intitulés en clair dans Hector.

[91] C'est la signification du symbole #.

```
# moyenne_en_maths DECIM 2
: MOYENNE Notes_en_mathématiques ;
```

Les variables calculées ainsi peuvent ensuite être utilisées comme n'importe quelle autre variable, y compris être réemployées dans d'autres formules. Chaque variable formulée conserve le texte de sa formule : elle peut ainsi être rejouée, c'est-à-dire calculée de nouveau, si des changements sont intervenus dans les données, comme des corrections ou des ajouts de sujets.

C'est aussi à travers les formules qu'est mis en œuvre le principe des bases de données relationnelles, qui permet par exemple dans une école d'échanger des informations et des calculs entre une table des élèves, une table des classes et une table des professeurs, chaque table étant contenue dans un corpus distinct, et reliée aux autres au moment du calcul d'*héritage* ou de *collecte*, ces mécanismes simples et robustes jouant le rôle des requêtes dans une base de données relationnelles classique.

La récupération des résultats dans un document

Partout dans Hector où sont affichés des résultats, l'utilisateur a la possibilité d'imprimer ces résultats ou de les envoyer dans un *document*.

L'impression immédiate est une tentation dont il faut se défendre sauf circonstances particulières, car outre qu'elle amène à gaspiller de l'encre et du papier, elle est de peu d'utilité dans la mesure où les résultats statistiques sont le plus souvent inclus dans un discours explicatif, descriptif ou démonstratif.

Jusqu'en 2008

Hector utilise donc, exclusivement jusqu'en 2008, la notion de document, qui revêt ici un sens spécial : le *document* dans Hector est un fichier de transport, d'un format de traitement de texte, dans lequel les résultats tabulaires, graphiques ou statistiques sont ajoutés chaque fois qu'on le demande. La démarche normale est ensuite *d'inclure* ce fichier dans le véritable document sur lequel on travaille, rapport, mémoire ou thèse. On peut se représenter le document de transport comme le caddy avec lequel on fait ses courses : il n'est qu'un moyen de stockage temporaire, et on ne fait pas ses courses avec son frigo.

Les textes et graphiques transportés par le document sont *vivants*, c'est-à-dire qu'il ne s'agit pas d'une simple copie inerte de l'image affichée, mais de véritables textes et de véritables graphiques qui peuvent être ensuite enrichis et retravaillés. Des options gouvernent la gestion des grands tableaux, la taille des images, etc.

Hector[2] propose quatre formats pour le fichier document ; le choix dépend essentiellement des autres outils qu'on emploie et notamment du traitement de texte utilisé :

- RTF, ou Rich Text File, qui est un format de transport issu de l'industrie de l'impression. Les graphiques y sont inclus dans le format EMF, ou Enhanced Metafile Format, format de graphique vectoriel qui a le défaut d'être la propriété de Microsoft ™ bien qu'il soit aussi lu par les traitements de texte du logiciel libre. Cette solution fonctionne bien si l'on travaille ses textes avec Word ®, en tous cas à la date où nous écrivons, car les formats des versions successives de ce logiciel peuvent changer.
- ODT, pour Open office Document Text. Ce format[93] est compatible avec le logiciel gratuit Open Office, et comporte ses propres instructions de graphique vectoriel. C'est le format recommandé si on n'a pas encore d'autres habitudes.

---

[92] Le mot clef DECIM fixe ici le nombre de chiffres après la virgule.

[93] Pour les curieux, il s'agit en fait d'un ensemble de fichiers XML comprimés dans une même archive, comme tout un chacun peut s'en assurer en changeant l'extension ODT en ZIP, avant de le

- TEX, ou plus précisément T_EX[94], format inventé par Donald Knuth, et qui est le favori des mathématiciens et informaticiens. Son utilisation nécessite un certain investissement intellectuel, mais ses usagers ne jurent que par lui. Le graphique vectoriel y est également inclus.
- HTML, pour HyperText Markup Langage, qui est le format des pages internet. Les graphiques y sont inclus au format SVG, ce qui peut poser problème car l'interprétation de ces graphiques nécessite un *plug-in*[95].

Depuis 2008, le copier-coller

Un mécanisme plus simple et plus rapide a été introduit après 2008 : le copier-coller direct depuis l'écran d'Hector vers un logiciel de traitement de texte simultanément ouvert, tel que Word de Microsoft ou le logiciel gratuit Libre Office (successeur d'Open Office).

Un simple clic ou une combinaison shift-clic, ctrl-clic ou alt-clic permet de copier séparément les graphiques, les textes et les tableaux statistiques, pour aller les coller dans le logiciel-cible. L'usage du « document » d'Hector n'est donc généralement plus utile, sauf dans le cas de très grands tableaux où Hector propose un découpage automatique, le rendu des tableaux complexes ou spéciaux (coloration des associations locales). Il est donc conservé pour ces cas particuliers.

Le copier-coller des graphiques fonctionne également vers de nombreux éditeurs graphiques vectoriels, car les graphiques d'Hector sont du dessin vectoriel « vivant » et donc éditable. Il faut néanmoins signaler que pour certains éditeurs graphiques gratuits le résultat peut-être altéré, avec parfois une disparition des traits fins ou une anamorphose indésirable. Il faut donc éventuellement procéder à des essais, et l'auteur d'Hector ne peut être tenu pour responsable de dysfonctionnements avec des logiciels inconnus de lui.

La version professionnelle-recherche d'Hector

Elle est en fait entièrement incluse dans la version de base[96], mais cachée. Elle ne devient accessible que si on actionne une clé et un mot de passe obtenus auprès de l'auteur. Ce dispositif présente l'immense avantage que l'auteur ne maintient, de son point de vue, qu'un seul logiciel, ce qui est déjà un gros travail, et que le point de téléchargement est unique.

Cette version professionnelle et de recherche ajoute cinq catégories d'éléments :

- Des éléments complémentaires dans l'interface de base
- Des éléments de compatibilité ascendante
- Des éléments statistiques avancés classiques
- Des éléments moins classiques voire carrément spéciaux
- Des éléments totalement originaux

Eléments complémentaires dans l'interface de base

Ces éléments apportent à l'interface de base des compléments utiles dans une approche professionnelle du logiciel, mais qui demandent un peu d'investissement intellectuel, au-delà de ce qu'il est de coutume d'attendre des apprenants auxquels est destinée la version de base.

Le cadre des éléments techniques avancés

Il apparaît dans la page d'accueil, et permet les opérations suivantes :

---

décomprimer (dézipper) le fichier ODT. Le format DOCX de Microsoft est construit sur les mêmes principes.

[94] Pour Tau Epsilon Khi, d'où la prononciation [tek].

[95] Un élément logiciel complémentaire qu'on télécharge le plus souvent gratuitement.

[96] Dont le caractère basique est quand même très relatif !

- Afficher des informations techniques détaillées sur le corpus
- Reprendre un tableau extérieur au format texte et en faire un corpus. Cette fonction permet l'analyse secondaire de tableaux croisés d'origine externe.
- Exporter comme corpus-fils : cette fonctionnalité, très utile dans certains cas de figure, est trop complexe pour être expliquée ici.

Le clonage des variables

Cette fonctionnalité de la gestion des variables permet, dans les corpus qui nécessitent un grand nombre de variables organisées en séries dont les noms se ressemblent à une variante près, de ne créer réellement avec tous ses paramètres qu'une série prototype, qu'on va ensuite reproduire automatiquement en indiquant seulement ce qui doit être changé dans l'intitulé. Cette manière de procéder sert non seulement à raccourcir une phase assez fastidieuse de la préparation du corpus, mais aussi à garantir la régularité du paramétrage des variables et de leur intitulé, régularité qui est par exemple indispensable à l'exportation comme corpus-fils.

Le plan de projection

Cette fonctionnalité de traitement permet de créer un plan graphique défini par un couple de variables numériques[97], appelées axes du plan. Dans ce plan, on peut projeter :

- Les sujets, ou individus statistiques,
    - soit comme un nuage de densité,
    - soit comme des symboles différents selon les catégories définies par une variable nominale ou ordinale
    - soit comme des points de couleurs différentes selon les catégories définies par une variable nominale ou ordinale
- Les variables.
    - Les numériques-calendaires sont représentées par un vecteur défini par leurs corrélations avec les axes.
    - Les ordinales, nominales et ordinales sont représentées par leurs étiquettes, positionnées aux coordonnées moyennes des sujets relevant de chaque catégorie. Optionnellement, la position est entourée d'une ellipse dont les pseudo-rayons sont les écart-types (affectés éventuellement d'un coefficient) de la distribution des positions des sujets de la catégorie selon les deux axes.
    - Optionnellement, les positions des ordinales sont reliées par un segment.

Diverses options sont proposées pour régler plus finement les modalités de représentation. Comme tout résultat graphique, le tracé du plan de projection peut être envoyé dans le document de transport ou copié-collé vers un traitement de texte.

Dans ce guide, le plan de projection est utilisé dans un exemple commenté à la fin de la rubrique « Une étude de cas de validation interne sur un test orthophonique ».

Eléments statistiques avancés classiques

Il s'agit de fonctionnalités de calcul factoriel qu'on retrouve dans la plupart des logiciels d'analyse de données[98], et qui sont utilisées depuis plusieurs dizaines d'années dans les

[97] Lesquelles peuvent, indifféremment, être des variables d'origine ou des variables issues d'axes factoriels (cf. éléments avancés classiques). Alors que la plupart des logiciels de statistiques mêlent la dimension de calcul et la dimension de représentation (on calcule une analyse factorielle *et* on l'affiche), Hector les sépare pour une plus grande versatilité (on calcule *ici* et on affiche *là*).

[98] Mais rarement dans la version de base au prix abordable, plutôt dans de coûteuses extensions.

recherches en Sciences Humaines. Les méthodes présentées ici ne sont pas les seules existantes[99], mais elles constituent un jeu raisonnablement complet.

L'analyse factorielle des correspondances

En abrégé, AFC, simple (deux variables) ou multiple (un nombre quelconque). Les fondements mathématiques sont exposés dans les ouvrages de statistiques ; on les résumera simplement ici en disant qu'il s'agit de réduire la complexité de tableaux de contingence à une série d'axes orthogonaux et d'inertie décroissante, dont on n'exploite que quelques-uns des premiers[100], le plus souvent à l'aide du plan de projection. L'AFC ne concerne que des variables catégorielles, à savoir les nominales, les ordinales et les logiques.

L'AFC produit à la demande le nombre d'axes demandés, en créant des variables porteuses d'intitulés formés initialement au moyen de la date et de l'heure, mais qu'il est conseillé de remplacer au plus vite par des intitulés porteurs de sens. Cette AFC qui crée des variables ne doit pas être confondue avec l'AFC simple illustrative, graphe du croisement de deux nominales, qui ne conserve pas de variables calculées.

L'analyse en composantes principales

Ou ACP. C'est le pendant de l'AFC, mais pour les variables numériques et calendaires ou encore logiques[101].

Elle peut être décrite comme l'analyse des multiples corrélations entre variables, et fournit comme l'AFC des variables numériques qui peuvent être réutilisées dans différents calculs et notamment comme axes du plan de projection.

La typologie

La fonctionnalité de typologie d'Hector opère sur des collections de type numérique ou assimilé (logique et calendaire), et cherche à constituer un découpage de la population étudiée en classes aussi homogènes que possible à l'interne et aussi diversifiées que possible à l'externe. La méthode s'appuie sur un algorithme de classification hiérarchique descendante obéissant aux règles suivantes :

- il commence par calculer le centre de gravité de la galaxie des sujets
- puis il regroupe les sujets identiques dans des groupes (il ne manquerait plus qu'ils fussent classés séparément)
- puis il classe les groupes de sujets (qu'on appellera ci-dessous sujets pour simplifier) par distance (euclidienne) croissante par rapport à ce centre
- puis il prend les deux premiers sujets de sa liste et décide d'en faire un bouquet (comme deux cerises)
- puis il prend le sujet suivant dans sa liste, et se demande s'il « vaudrait mieux »
    - l'accrocher au sujet de gauche
    - l'accrocher au sujet de droite
    - faire un nouveau bouquet avec d'un côté l'ancien bouquet et de l'autre le sujet en cours de classement
- puis il se décide, et fait ça pour chacun des sujets jusqu'à épuisement de la liste

A la fin, il dispose d'une arborescence complète, quelque chose comme un chou-fleur avec un gros trognon qui est l'ensemble des sujets, et des bouquets qui se subdivisent jusqu'à n'être plus, aux extrémités, que des sujets individuels ou des paquets de sujets identiques.

---

[99] Les méthodes statistiques sont innombrables, et de nouvelles apparaissent chaque jour.

[100] Et souvent seulement les deux premiers, s'ils rassemblent une proportion raisonnable de l'inertie totale (60 ou 75 % par exemple), permettant de négliger le reste, considéré comme du *bruit*.

[101] Traitées en cette occasion comme des numériques aux valeurs {0,1}.

L'usager décide ensuite du nombre de classes dont il désire disposer, et peut choisir plusieurs options de critères de coupure. L'idéal d'une typologie étant un nombre de classes assez élevé pour rendre compte des différences, mais en même temps des contrastes assez caractérisés pour pouvoir nommer les classes obtenues de manière intelligible pour le lecteur du texte final, la mise au point d'une typologie est une recherche de compromis entre besoins contradictoires.

Quand les choix sont arrêtés, la typologie génère à la demande une nouvelle variable nominale dont l'intitulé est fabriqué à l'aide de la date et de l'heure, et dont les étiquettes sont des textes stéréotypés du genre cl.1, cl.2 etc. Il convient évidemment de donner au plus vite un autre nom et d'autres étiquettes, en s'aidant de l'analyse des profils de chacune des classes.

La régression multiple

Cette fonctionnalité, introduite en Octobre 2008, est une des variétés de l'analyse des corrélations. Celle-ci est au principe de plusieurs méthodes, concurrentes et en évolution constante, qui visent toutes, d'une manière ou d'une autre, à faire apparaître des variables *latentes*, c'est-à-dire des grandeurs qui n'ont pas été observées directement mais qu'on suppose sous-jacentes aux grandeurs observées, et capables dans certains cas de les *expliquer* ou de les *prédire*. Il s'agit donc d'une première étape vers les diverses approches relatives aux variables latentes, aux pistes causales et aux équations structurelles, qui constituent le principal axe de développement actuel d'Hector.

Fondamentalement, la démarche consiste à optimiser les coefficients de régression b_i dans une équation telle que :

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots + b_px_p$$

où $\hat{y}$ est un estimateur de la variable dépendante y (aussi appelée *critère*) calculé comme une combinaison linéaire des variable indépendantes x_i (aussi appelées *prédicteurs*), de sorte que la corrélation entre y et $\hat{y}$ soit maximale.

Les modèles d'équations structurales

Les Modéles d'Equations Structurales, plus souvent appelées SEM (Structural equation model), et encore plus souvent confondues avec LISREL ©, qui en est la version commerciale la plus diffusée, portent au Québec le nom d'analyses cheminatoires, ou acheminatoires.

Parfois appelée aussi Analyse Factorielle Confirmatoire, l'approche SEM se donne pour objet d'évaluer la pertinence de modèles de relations entre variables, parfois complexes, mais dont les plus simples se ramènent à l'hypothèse que les valeurs des variables observées peuvent être décrites comme des combinaisons linéaires des variables latentes.

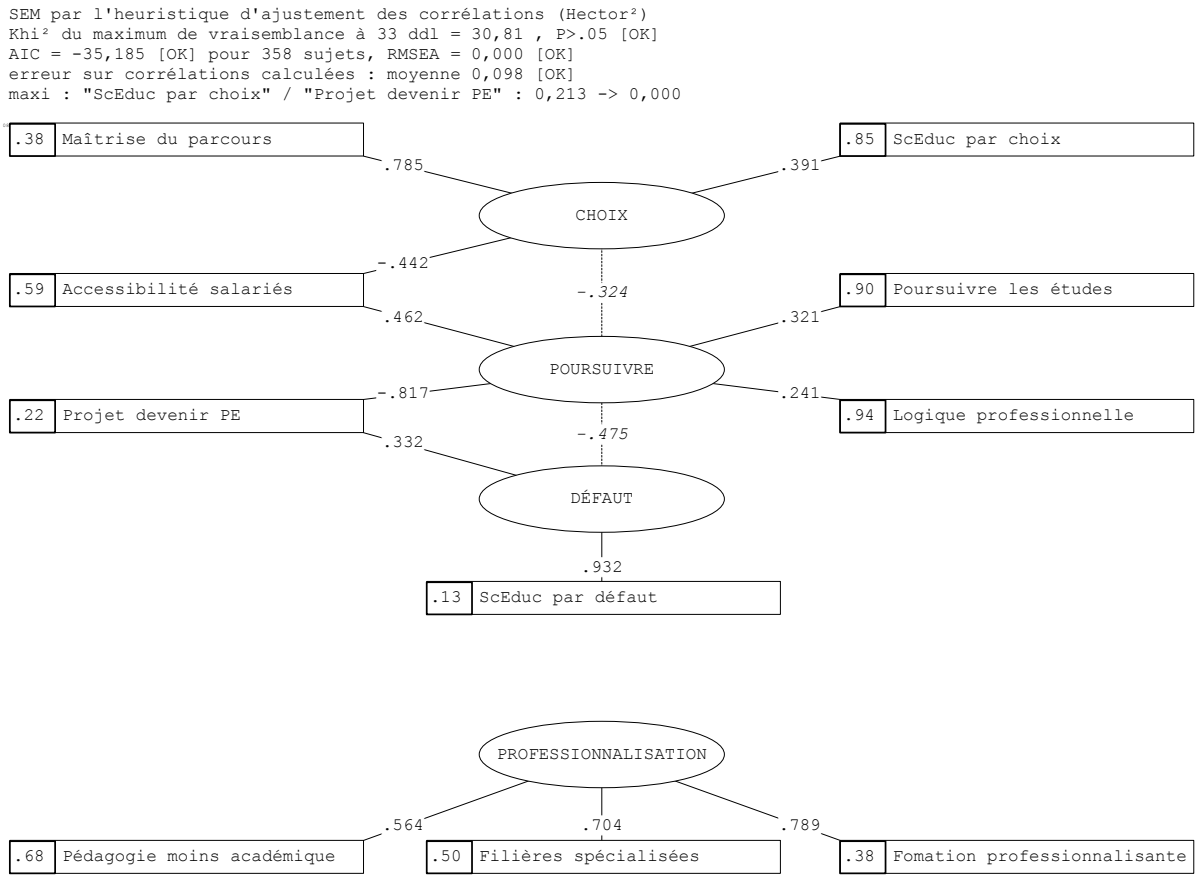
On fournit donc au moteur SEM la grille des relations dont on suppose l'existence, par exemple en application d'un modèle théorique, et le moteur se charge, sur la base de l'analyse des corrélations entre variables observées :

- D'estimer la valeur des pondérations des variables latentes dans les équations structurales
- D'estimer la vraisemblance du modèle, c'est-à-dire à quel point les valeurs des corrélations qu'il « prédit » pour les variables observées sont proches des valeurs réellement connues.

La vocation affichée des SEM est donc la vérification de la pertinence de modèles élaborés par ailleurs, en conformité avec des hypothèses appuyées sur une théorie. Dans l'usage concret, la facilité interactive avec laquelle on peut tester des modèles variés peut leur conférer un usage exploratoire.

L'implémentation des SEM dans Hector utilise un Heuristique original d'ajustement des corrélations entre variables observées et variables latentes.

L'approche SEM est actuellement très populaire en Sciences Humaines, à la fois par la qualité de synthèse qu'elle apporte à l'analyse des corrélations et la clarté des graphiques qu'elle engendre, comme dans l'exemple suivant :



La mise en ordre du graphique est assurée dans Hector par un éditeur graphique intégré. L'exemple ci-dessus est extrait du manuel Hector 2 approches factorielles 2011.pdf, où il est commenté en détail.

Eléments moins classiques voire carrément spéciaux

On range dans cette rubrique les spécialités, c'est-à-dire les traitements pas nécessairement orthodoxes dont l'auteur a éprouvé le besoin pour une recherche particulière, ou qui lui ont été demandés par un partenaire dans un but particulier. Chacun des éléments de cette rubrique a une histoire, qui est rappelée dans la documentation, et certains aspects de ces fonctionnalités répondent à des besoins précis, mais ont été conçus et documentés de manière à pouvoir satisfaire des besoins similaires, et inclus à ce titre dans la version professionnelle-recherche.

Dichotomies

Il s'agit d'un outil rugueux mais puissant d'exploration de données.

Il n'existe pas de test statistique universel permettant de comparer entre elles des variables de type quelconque.

Toutefois, il existe un test polysémique pour les cas où chaque variable est réduite à deux valeurs distinctes, autrement dit à une *dichotomie*. Il s'agit du φ, ou phi, qui s'interprète comme un coefficient de corrélation[102], mais entretient des liens avec le khi², puisque

[102] Valeurs comprises entre –1 et +1, 0 correspondant à l'indépendance.

$khi^2/N = \varphi^2$. Bref un véritable couteau suisse, pourvu que les variables soient dichotomiques.

Or, de toute variable on peut extraire une ou plusieurs dichotomies :

- Pour une variable impliquant un ordre, comme une calendaire, une numérique ou un ordinale, tout point de coupure dans la série est susceptible de créer une dichotomie[103].
- Pour une variable non-ordonnée, la nominale, les dichotomies sont des oppositions sur des partitions complémentaires des catégories. Ainsi une variable nominale à quatre valeurs distinctes a, b, c, d admet sept dichotomies[104] : a / bcd , b / acd , c / abd, d / abc, ab / cd, ac / bd, ad / bc.
- Enfin les logiques sont des dichotomies par nature

La fonction de recherche des dichotomies calcule donc *toutes* les dichotomies sur les variables citées directement ou par leurs collections, et s'efforce de repérer celles qui fournissent les valeurs les plus élevées, en valeur absolue, de φ.

Différents instruments de paramétrage permettent de l'imiter l'affichage à des valeurs significatives à un certain seuil, ou à un certain nombre de valeurs par couple de variables.

En dépit de la puissance de la fonction, qui permet de repérer d'un simple coup d'œil où vont se trouver les relations intéressantes[105], l'usager est invité à la sobriété, notamment avec les variables nominales dont le nombre de dichotomies croît de manière exponentielle avec le nombre de valeurs distinctes, sachant que le nombre de croisements élémentaires est le *produit* du nombre des dichotomies engendrées par les variables.

Collections parallèles

Des collections parallèles sont des collections de variables du même type, et qui se correspondent de collection à collection dans le même ordre, selon un lien sémantique qui pourrait être, par exemple, que la $n^{ème}$ variable de la première collection est la mesure *avant* de quelque chose qui est mesuré *après* par la $n^{ème}$ variable de la seconde collection. Ce peut donc être une mesure d'évolution ou de progrès entre un temps t0 et un temps t1, ou de part et d'autre d'un apprentissage, mais ce lien n'est qu'un exemple.

L'important est que les variables soient numériques, calendaires ou logiques, en nombre égal dans chacune des collections invoquées (qui ne sont pas limitées à deux), le parallélisme sémantique étant entièrement du ressort de l'usager.

La fonction calcule, pour chaque couple de valeurs, le coefficient de corrélation r de Bravais-Pearson et le |t| de Student sur échantillon appareillés[106].

Différentes options de présentation et de choix des seuils de significativité sont disponibles, et l'interprétation dépend évidemment de la nature des données, mais,on peut énoncer les principes suivants : le coefficient de corrélation r mesurant la *cohérence* entre les mesures et le |t| de Student mesurant la significativité de la différence des

---

[103] si la variable comporte n valeurs distinctes, le nombre de dichotomies possibles est donc de n-1.

[104] Le nombre des parties d'une ensemble de cardinal n est 2^n. Cependant le nombre des dichotomies est deux fois moindre, car à chaque partie de l'ensemble correspond son complémentaire, et une dichotomie a / bcd n'est pas distincte de la dichotomie bcd / a. On en est donc à 2^{n-1} ; mais reste à considérer une dichotomie triviale : rien / abcd, ou abcd /rien. Cette dichotomie théorique est donc dépourvu d'intérêt pratique. Le nombre de dichotomies utiles sur la base de n classes est donc de 2^{n-1} –1.

[105] Parce qu'un coefficient φ élevé sur les dichotomies entre deux variables est prometteur de valeurs élevées pour les statistiques de croisement appropriées au type initial des variables. Le phi est donc moins intéressant en soi que comme signal ou indice de *où chercher*.

[106] Autrement dit, deux mesures différentes sur les mêmes sujets, par opposition au |t| de Student sur échantillons indépendants, mesures identiques sur deux groupes distincts de sujets.

moyennes, en termes de progression ou d régression, mais en tous cas de *changement*, quatre cas de figure peuvent se présenter, en fonction des seuils choisis :

- |t| et r sont significatifs : il y a changement, et il s'opère de manière cohérente pour l'ensemble des sujets ; le nuage des valeurs s'est déplacé sans se déformer
- |t| est significatif, mais pas r : il y a eu globalement un changement significatif, mais sans cohérence ; le nuage des valeurs s'est déplacé, mais en se déformant, des points ont changé de place
- |t| n'est pas significatif, mais r l'est : il n'y a pas eu de changement d'ensemble, et la cohérence a été relativement maintenue ; le nuage est resté sur place sans se déformer
- ni |t|, ni r ne sont significatifs : il n'y a pas eu de changement global, mais des mouvements internes désordonnés, browniens ; le nuage est resté globalement sur place, mais en se déformant

Les explications données ci-dessus se veulent analogiques, mais on comprendra l'intérêt de telles fonctions pour la mesure de l'efficacité des formations, l'évolution des représentations ou la fidélité des tests.

Méta-formule

Une méta-formule n'est rien d'autre qu'une formule de calcul de nouvelles variables qui s'applique à une méta-collection, c'est-à-dire à un ensemble de collections. Autrement dit, la méta-collection est un étage de plus dans la classification des objets dans Hector.

Le mode d'emploi, tout simplet, consiste à écrire une formule impliquant une collection[107], et à remplacer le nom de la collection par le joker ???, puis à co-sélectionner plusieurs collections, avant de demander l'exécution.

Le but du jeu, sans prétentions, est de gagner du temps dans des opérations répétitives et fastidieuses. Si on a par exemple préparé plusieurs collections d'items correspondant chacune à des compétences distinctes, on calculera d'un seul les scores de totalisation pour chacune de ces compétences.

LPE / Consensus

Il s'agit moins ici d'un outil d'élaboration statistique que d'un instrument de présentation de résultats.

Il s'agit d'obtenir le genre de résultat qui figure dans le graphique ci-après. Au départ on dispose d'une série d'indicateurs binaires[108] qui ont été relevés sur une population divisée en plusieurs catégories[109].

On a la double intention de montrer comment ces indicateurs s'ordonnent du plus anodin au plus sensible, et comment les différentes catégories se situent au regard de ces indicateurs, permettant de repérer du premier coup d'œil quelle catégorie était susceptible de poser un problème particulier : une ligne relie les positions moyennes de l'ensemble des groupes sur les différents indicateurs, et chaque groupe est repéré par un point à gauche, à droite ou sur la ligne, une position très éloignée de la ligne réclamant une attention particulière.

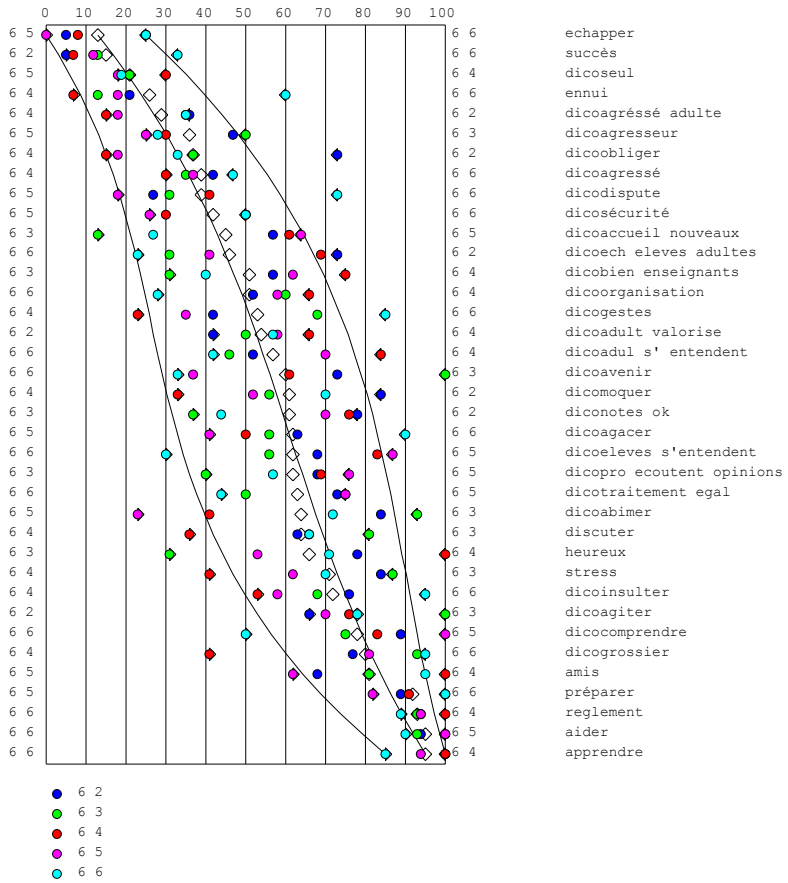
Au départ, les utilisateurs de cette technique l'avaient baptisé cette ligne LPE, pour ligne de partage des eaux, et le travail de représentation était effectué, très fastidieusement, à l'éditeur graphique.

---

[107] Formule qu'il vaut mieux avoir testé auparavant en individuel.

[108] Obtenus dans le cas historique à partir de valuateurs de représentations, mais ils pourraient avoir n'importe quelle autre origine.

[109] Historiquement, des classes dans un collège.



La fonction LPE n'apporte pas grand chose conceptuellement à cette démarche, mais permet en revanche de tester très rapidement une grande variété de paramétrages.

La fonction Consensus, complémentaire de la précédente, a été ajoutée spontanément par l'auteur.

Sur la même collection logique que précédemment, les opérations de consensus posent la question : existe-t-il un ordre des variables pour lequel les sujets seraient à peu près d'accord ? Cette question est importante puisqu'elle peut permettre d'élaborer un ordre canonique permanent (et non pas calculé à chaque fois) des variables, en éliminant celles qui ne font décidément pas consensus.

Les seuils d'accord utilisés sont évidemment paramétrables. Le but du jeu est de sélectionner un ensemble de variables dont les sujets s'accordent à considérer qu'ils constituent, à peu de variation près, une suite implicative logique. L'ensemble peut ensuite être réinjecté dans la partie LPE.

Quoique développé dans le cadre d'une analyse de représentations, ce modèle pourrait trouver son utilité en didactique comme en construction d'épreuves et de tests.

Connectivité

Le but de la démarche est le suivant : quand on a recueilli un grand nombre d'indicateurs d'éléments non directement tangibles tels que des compétences ou des attitudes, comment peut-on regrouper au mieux ces indicateurs pour restituer avec la plus grande force possible des variables latentes qui se rapprocheraient de la mesure inaccessible de ces éléments hypothétiques ?

Par connectivité, on entend donc la possibilité de connecter les items entre eux, pour former des agrégats d'items à la fois cohérents (critère formel) et intelligibles (critère sémantique), tant il est vrai qu'il ne sert à rien d'exhiber des entités d'origine mathématique dont on ne peut débattre, parce qu'on n'est pas en mesure de décrire ce qu'elles représentent.

C'est en quelque sorte une extension et une généralisation de la méthode d'arborescence des parentés dans la validation des tests. Elle s'applique cependant à de plus vastes ensembles de variables, pour lesquels l'arborescence graphique devient rapidement illisible.

Trois familles d'algorithmes sont proposées. Avec leurs variantes, elles proposent sept manières différentes et paramétrables d'extraire d'une grande liste d'items un certain nombre de groupes intéressants.

Ces trois familles reçoivent les noms poétiques d'algorithme du chou-fleur, d'algorithme de la boule de neige et d'algorithme de l'oignon.

Les algorithmes du chou-fleur consistent d'abord à organiser les items en petits bouquets, puis en bouquets de bouquets de plus en plus gros, jusqu'à la tête de chou fleur, qui est le bouquet suprême. Dans un second temps, on se propose de couper dans le chou-fleur, plus ou moins loin de son cœur, pour séparer les bouquets, qui seront les groupe d'items. Plus on coupe près du cœur, plus les bouquets sont gros et moins ils sont nombreux, et réciproquement. Ce n'est pas propre à l'analyse des items, ça se retrouve aussi dans la construction des typologies[110]. Les différences entre variantes reposent sur différentes manières de considérer à quel point un item est proche d'un bouquet.

Les algorithmes de la boule de neige consistent à partir d'un noyau, qui est toujours constitué des deux variables les plus ressemblantes, et de chercher parmi les items non encore rangés lequel pourrait rejoindre la boule de neige et la rendre encore plus jolie, entendez par là encore plus cohérente ; on continue jusqu'à ce qu'aucun item ne mérite de rejoindre la belle boule ; alors on la range sur le côté, et on recommence avec les items en vrac qui restent, jusqu'à ce qu'il n'y ait plus d'items qui se ressemblent assez pour faire une boule.

L'algorithme de l'oignon est un peu l'inverse des précédents : on part de l'ensemble de tous les items, et on épluche les items les moins bien assortis à l'ensemble général, jusqu'à ce qu'on ne puisse plus rien enlever sans nuire à la beauté de l'oignon. Alors on met l'oignon de côté, on ramasse les pluches, on les rassemble et on essaye de faire un nouvel oignon.

Ces trois algorithmes sont susceptibles de fournir des résultats différents selon la nature des données et les nombreuses possibilité de paramétrage disponibles. Chacun a ses qualités et ses défauts, et l'usager doit se livrer à des essais, mais aussi faire usage de son sens critique. En fait, ce n'est pas au logiciel de décider : il est un auxiliaire, et on peut utiliser plusieurs outils successivement, l'important étant de constituer des agrégats d'items d'ont on va pouvoir vérifier la robustesse au moyen d'outils plus classiques comme les tests de cohérence item-test et l'alpha de Cronbach. On notera cependant que les divers algorithmes fournissent parfois une certaine régularité dans les regroupements d'items, et on peut alors parler de *formes fortes*, dont la cohérence est certaine.

Distances

Il s'agit de variables numériques, et la démarche repose ici sur l'analyse de la matrice des corrélations. Sur la base de la corrélation de Bravais-Pearson de valeur r entre deux variables, avec $-1 \leq r \leq 1$, on cherche une mesure de dissimilarité[111] δ qui possède les caractéristiques suivantes : $\delta ii = 0$ est le cas d'identité, $\delta ij \geq 0$ affirme la positivité et $\delta ij = \delta ji$ exprime la symétrie. L'identité $\delta ii = 0$ signifie que la dissimilarité entre un objet et lui-même est nulle, puisqu'il est intégralement semblable à lui-même. La positivité $\delta ij \geq 0$ indique qu'une dissimilarité ne saurait être négative. Si elle est nulle, elle indique la

[110] Démarche commune avec les typologies, précédemment exposées. Les typologies sont des analyses de connectivité sur les sujets, et non, comme ici, sur les variables.

[111] La dissimilarité est l'antonyme de la ressemblance, au point que beaucoup de dissimilarités sont construites en comptant entre les objets étudiés les points de ressemblance, puis en soustrayant le résultat obtenu du maximum théoriquement possible.

ressemblance maximale. La symétrie $\delta ij = \delta ji$ indique que la distance entre un objet i et un objet j est la même que dans l'autre sens.

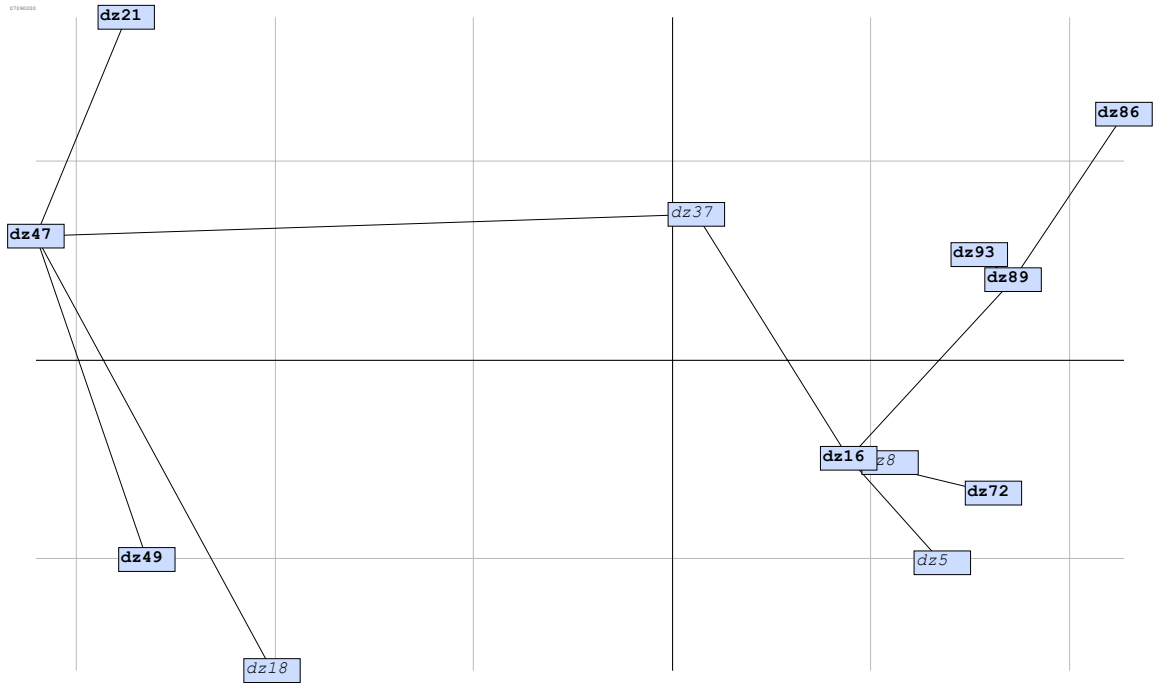
On peut constater que toutes les distances physiques sont des dissimilarités, mais que l'inverse n'est pas vrai : une condition supplémentaire pour qu'une dissimilarité soit une distance est l'inégalité triangulaire : $\delta ij \leq \delta ik + \delta jk$

Cette dernière formule exprime qu'on peut tracer un triangle dans un plan avec ces trois points, ou encore que c'est toujours plus court et en tous cas pas plus long d'aller tout droit que de faire un détour. C'est évident dans le domaine physique, mais pas forcément en mathématiques.

Or on sait analyser factoriellement un tableau de distances, mais pas un tableau de dissimilarités. On dispose heureusement d'un moyen de corriger un tableau de dissimilarités en distances en ajoutant à chaque distance, sauf à celles qui sont nulles et doivent le rester, la plus petite quantité nécessaire pour que l'inégalité triangulaire soit réalisée pour tous les trios d'objets. Cette transformation, qui conserve l'ordre des dissimilarités, ne pose pas de problèmes particuliers ensuite pour l'interprétation.

L'analyse d'un tableau de distances consiste à projeter l'ensemble des n objets dans un espace à n-1 dimensions respectant intégralement les distances, puis à extraire successivement des dimensions mutuellement orthogonales et d'inertie décroissante, comme dans toute analyse factorielle. On obtient ainsi les coordonnées de chaque objet dans chaque dimension, et on utilise généralement les deux premières (et les plus importantes) comme coordonnées dans un plan illustratif.

Outre la représentation dans le plan des objets analysés selon leurs distances, Hector propose également la superposition de différentes sortes de graphes :



Ici un *arbre minimum*, mais d'autres graphes sont disponibles, comme les graphes au *seuil de la moyenne* et au *seuil de connexité*, dont la description dépasse un peu le cadre de ce guide.

Eléments totalement originaux

L'auteur ne désespère pas de trouver encore quelques bonnes idées dans le traitement de données, mais il considère que la meilleure qu'il ait eue jusqu présent est incontestablement celle de l'analyse des séquences.

Une *séquence* est définie ici comme une série d'*états*. Les états sont pris dans un *vocabulaire des états*, ensemble fini et explicite des états possibles dans l'univers étudié.

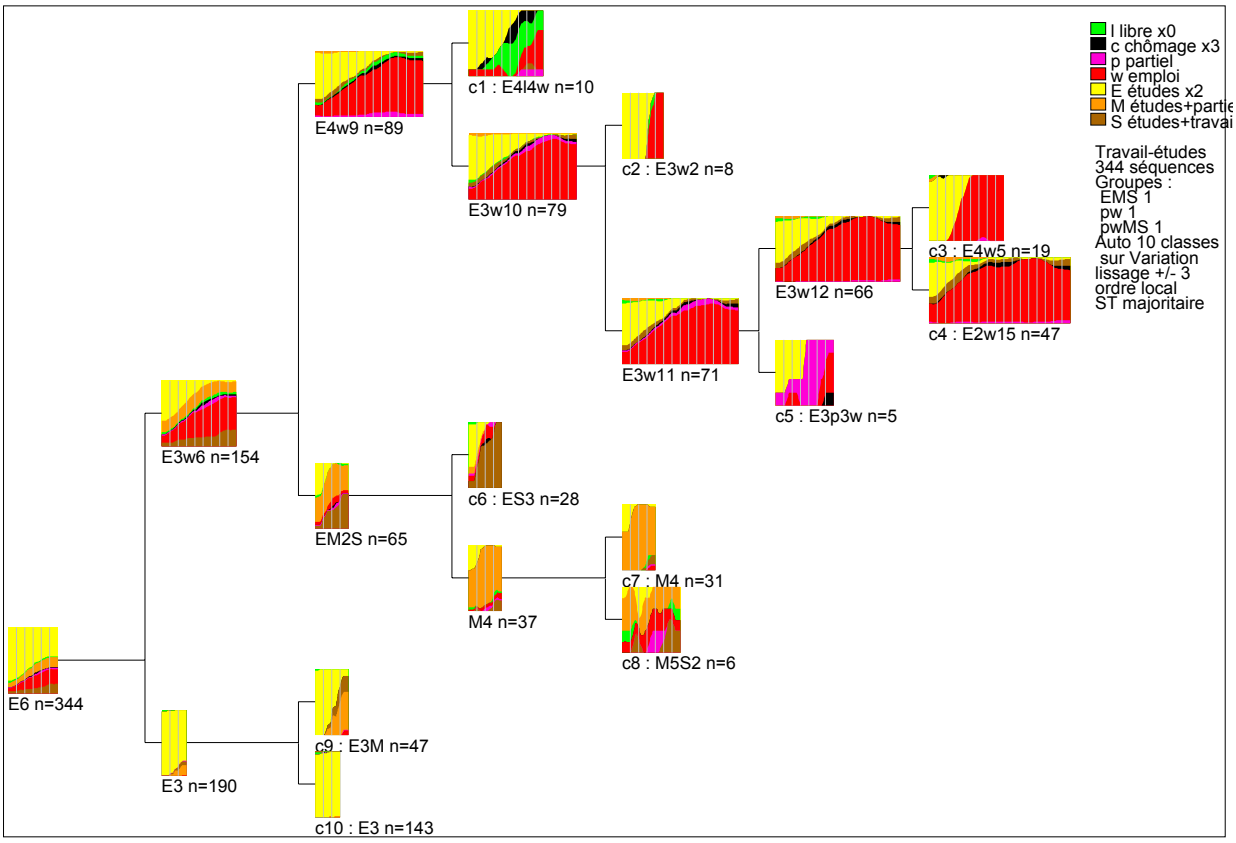
Un état peut apparaître une fois, plusieurs fois, aucune fois dans une séquence donnée. La longueur totale d'une séquence (le nombre des états successifs qu'elle contient) est quelconque.

Les états sont représentés par des symboles (essentiellement des lettres). Une séquence unitaire est donc représentée par une suite ininterrompue (sans espace ni caractère spécial du genre tabulation ou saut de ligne) de symboles. Cette séquence peut être codée dans le corpus comme une variable de type « texte libre ».

Un ensemble de séquences est constitué de plusieurs séquences, chacune d'elle étant associée à un individu statistique différent.

Le caractère successif des états d'une séquence a une portée très abstraite. Les données en forme de séquences peuvent renvoyer à des éléments très variés de la réalité : elles sont utilisables dès que le phénomène étudié peut être caractérisé par une succession d'entités prises dans un ensemble fermé, chaque entité étant caractéristique d'une étape ou période discrète ; la succession elle-même peut renvoyer à l'écoulement du temps, mais aussi à d'autres concepts (conséquence, filiation, place dans un classement…).

Sans entrer dans le détail de cette technique très riche dans ses possibilités d'analyse et d'interprétation de données jusque là inutilisable, on souhaite simplement ici donner envie d'en savoir plus, en montrant un exemple de ce que produit l'analyse des séquences :



Les données prises ici pour l'illustration de la technique proviennent d'une enquête auprès de plusieurs centaines d'étudiants en licence de sciences de l'éducation, au cours de l'année 2001. Les éléments faisant l'objet d'une analyse de séquence sont ici les différentes situations connues par les individus par rapport l'emploi, aux études ou à une combinaison des deux, entre la fin de leurs études secondaires et le début de l'année universitaire de l'enquête.

Que représente le schéma ? C'est le résultat d'une classification automatique des séquences selon la méthode de la Classification Hiérarchique Descendante, avec coupure

automatique à dix classes sur le critère de la variation, avec toutes les options graphiques : ce sont les paramètres par défaut[112].

La partie droite de l'image contient la Légende générale, c'est à dire la liste des états connus, avec leur couleur, leur légende particulière et leur pondération, s'ils en ont une. En dessous apparaissent des éléments d'information sur les options utilisées, ainsi que la liste des groupes d'états et leurs pondérations, s'ils en existe.

La partie gauche de l'image représente l'arborescence des classes obtenues par l'analyse.

Chaque vignette colorée résume la composition d'une classe : chaque tranche verticale représente une période, et la vignette comporte autant de périodes que les séquences qui constituent la classe en avaient en moyenne.

Dans chaque tranche, l'épaisseur d'une veine d'une couleur donnée correspond à la *densité de probabilité* de l'état correspondant dans cette période.

En dessous de chaque vignette, un texte décrit autrement le contenu de la classe : il fournit la *séquence typique* (ou S-T) de la classe, c'est-à-dire la série des états majoritaires dans les périodes successives. Si un état apparaît plus d'une fois, il est noté par son nom, avec le nombre de répétitions : E3 est une abréviation pour EEE. Cette séquence typique n'apparaît que s'il y a assez de place pour elle compte tenu de la taille de caractère actuellement utilisée (que l'on peut modifier). En revanche, apparaissent toujours en fin de texte l'effectif (n=143), et, en début de texte, la lettre c suivie d'un numéro de classe, s'il s'agit d'une *classe finale*, résultat du découpage.

Les classes qui ne portent pas de numéro sont les sur-classes : elles montrent comment l'ensemble des données se décompose, à partir de la sur-classe totale qui est représentée par la vignette de gauche. En fait, il s'agit d'une arborescence un peu particulière, puisque la racine est à gauche, que l'arbre est couché sur le côté, que les branches ont horizontales et que les feuilles (les classes finales) sont à droite.

La plupart des autres commandes de l'Analyse de séquences vont avoir pour but de travailler cette représentation et l'analyse sous-jacente dans le sens du but recherché, à savoir la meilleure typologie possible, celle qui allie dans le compromis le plus efficace un nombre raisonnable de classes finales et la possibilité de décrire et commenter celles-ci.

Comme cet objectif ne peut être probablement atteint qu'après un certain nombre d'essais et d'erreurs, les fonctions de Séquences sont conçues pour faciliter au maximum une démarche interactive.

D'autres éléments statistiques et de nombreux éléments de paramétrage accompagnent cette technique très prometteuse.

Dans le domaine de l'orthophonie, un domaine d'application pourrait être trouvé dans le codage de séries de productions verbales, ou de phases dans une stratégie de résolution de problème, ou encore dans l'analyse d'interactions orales, bref dans toute situation où les données sont caractérisée par le fait qu'elles sont *successives* et en nombre *indéterminé à l'avance*.

Comment évoluera Hector ?

Redisons-le avec Pierre Dac, la prédiction est un art difficile, notamment en ce qui concerne l'avenir. Cependant le passé d'Hector donne une certaine idée de son avenir. Tant qu'Hector aura des usagers qui apporteront des problèmes et des besoins, Hector continuera à se développer en ajoutant des solutions et des ressources. C'est en tous cas la direction dans laquelle l'auteur s'emploiera à poursuivre son activité, tant qu'il en aura la capacité et l'énergie.

[112] La signification de ces termes doit évidemment s'éclairer à la lecture du manuel

Annexe 3 : Cours de statistique élémentaire

Ceci est la réécriture, en Novembre 2011, d'éléments de cours de statistiques élaborés aux alentours de 2005 à l'intention d'étudiants de troisième année de licence de Sciences de l'Education et d'étudiants de troisième année d'Orthophonie. Ce cours repose sur un effort considérable de simplification – on pourra la trouver abusive - des modes d'exposition classiques des principes statistiques. L'idée de base est que les données qu'on étudie relèvent d'un nombre limité de types, fondamentalement, deux et pas plus, et que le type des données gouverne les opérations qu'on peut légitimement leur appliquer, ce qui donne, en statistique inductive, trois genres de tris simultanés de deux variables, ou croisements, et pas d'avantage, et que tout le reste en découle.

Cette prétention un peu folle de ramener toute la statistique descriptive et inductive à cinq opérations fondamentales, deux pour la descriptive et trois pour l'inductive, était née de la nécessité de proposer dans un temps sévèrement limité une initiation statistique utile aux stagiaires du DUERFO, Diplôme d'Université pour l'Enseignement, la Recherche et la Formation en Orthophonie, organisé par l'Institut d'Orthophonie de Lille II pour les praticien-ne-s qui souhaitaient s'engager dans des activités de formation et de recherche. C'était le début d'une coopération avec le monde de l'orthophonie qui dure encore. Parallèlement, ce fut aussi l'occasion de faire progresser mes outils d'enseignement et de recherche, et notamment de développer Hector, qui remplaça en 2004 le Nestor de 2001 et tous ceux qui l'avaient précédé. C'est avec ce logiciel Hector que sont créés les tableaux et graphiques qui illustrent ici la démarche, mais celle-ci est indépendante de l'outil employé, même si, bien sûr, j'ai tendance à considérer celui-ci comme bien meilleur que maint autre.

Le type des variables

Une variable, c'est quelque chose qui a au plus une valeur pour un individu statistique (un sujet, une observation), mais qui peut avoir des valeurs différentes pour différents individus.

Des valeurs différentes, mais pas n'importe quelle valeur. La taille d'un individu ne peut pas avoir la valeur « bleu », et sa nationalité ne peut pas être « un mètre cinquante ». Les variables sont dotées d'un type, qui gouverne les valeurs qu'elles peuvent prendre, et ce type détermine strictement et sans recours le genre d'opérations qui sont autorisées sur les données : on ne peut calculer une moyenne que sur des nombres, pas sur des nationalités…

Il s'agit donc d'une notion de la plus haute importance, qui conditionne la compréhension de tout ce qui suit.

Mesure et catégorie

C'est la première grande distinction entre types de variables.

Une variable qui provient d'une mesure (qu'on peut donc dire aussi métrique) est exprimée par des nombres (et on la dira aussi *numérique*). Ces nombres peuvent être entiers, réels, positifs, négatifs : ce sont des détails secondaires. L'important est que ces nombres proviennent, soit d'un comptage d'objets (combien le sujet possède-t-il de moutons ?), soit de la comparaison d'une caractéristique du sujet avec un étalon (quelle est la taille du sujet). On dit qu'on est en présence d'une échelle de mesure. La question de savoir si cette échelle à un zéro correspondant à une nullité dans le monde réel (le nombre de cheveux sur votre tête) ou si c'est une échelle à zéro conventionnel comme la température est sans conséquences pratiques pour ce qui nous intéresse.

L'autre grand type de variable est la catégorie, c'est-à-dire le sous-ensemble dans lequel on peut classer le sujet d'un certain point de vue, comme la nationalité, le sexe, la

couleur des yeux, le style vestimentaire, la préférence partisane. Ce qui caractérise les valeurs des variables de ce type est qu'elle sont repérées par des *textes* arbitraires, qu'on appelle aussi *étiquettes*. Ces textes ne sont bien sûr pas arbitraires pour le lecteur, qui préfère pour le sexe utiliser {garçon} et {fille} plutôt que {plouf} et {tagada}, mais ce l'est totalement pour le logiciel de calcul, qui n'a aucune espèce de conscience du monde réel et donc ignore la sémantique, alors que ce ne serait nullement indifférent d'utiliser 3 plutôt que 2 dans une variable numérique.

Donc, à ce stade du propos, on oppose des variables numériques (dites aussi quantitatives) aux variables texte (dites aussi qualitatives).

Avec une variable numérique, on peut dire des valeurs x et y concernant deux sujets que $x=y$, ou que $x\neq y$, ou que $x\geq y$ et même que $x-y=k$ (on peut calculer une différence).

Avec une variable texte, on peut juste dire $x=y$ (les deux sujets appartiennent à la même catégorie) ou $x\neq y$ (les deux sujets n'appartiennent pas à la même catégorie), mais pas $x<y$ (le patois picard n'est pas inférieur au patois normand, quoiqu'en pensent les locuteurs de ce dernier idiome).

Métriques

Un nombre est un nombre, mais certains nombres ont des usages particuliers. Ainsi le nombre qui code une date comme un certain nombre de jours passés depuis un jour choisi comme le début peut-il s'exprimer effectivement avec un nombre de jours en vrac, ou comme une date : 11/02/2005. Dans ce cas, on parlera de type *calendaire*, et on réservera le type *numérique* aux nombres normaux.

Textes

Là aussi, une nuance importante peut être introduite. Le patois picard n'est peut-être pas inférieur au patois normand, mais si je suis étudiant, je sais parfaitement que la mention Très Bien est supérieure à la mention Bien. Quelle sorte de type de variable est-ce donc là ? Elle s'exprime avec des étiquettes arbitraires, mais admet un *ordre*. On l'appellera variable de type *ordinal*.

Attention ! Une variable ayant pour valeurs possibles {pas du tout} {un peu} {beaucoup} {passionnément} {à la folie} est ordinale, mais la plus traditionnelle série {un peu} {beaucoup} {passionnément} {à la folie} {pas du tout} ne l'est pas : l'ordre doit être significatif et immuable. On notera qu'on peut estimer que $x>y$ pour une ordinale (ça veut dire que x est *après* y dans l'ordre), mais pas calculer $x-y=k$: l'écart entre {un peu} et {beaucoup} n'est en aucune façon comparable à l'écart entre {pas du tout} et {un peu} ; les mathématiciens diraient que ce n'est pas une échelle d'intervalles. Nous nous contenterons de remarquer qu'on ne peut pas calculer une moyenne sur ce genre de variable.

En conclusion pour les variables dont les valeurs sont des étiquettes (des textes), on distinguera le type *ordinal* si l'ordre des étiquettes est fixe et possède une signification, le type *nominal* si ce n'est pas le cas (nationalités, professions).

Vrai ou Faux

Reste le cas très particulier des variables qui ne peuvent prendre leur valeur que parmi DEUX possibilités (on les dit parfois, pour cette raison, *binaires*). C'est le cas des oppositions OUI/NON, VRAI/FAUX, 0/1, absent/présent …

Ce type très simple a beaucoup de vertus : il a les qualités des variables texte, avec comme étiquettes {vrai} et {faux}, mais il est évidemment ordinal, puisque deux objets sont toujours classés l'un par rapport à l'autre ; enfin, il peut être considéré comme numérique avec les valeurs 0 et 1, et on peut calculer des moyennes qui sont alors des probabilités.

Surtout, ce type admet les opérations de la logique des propositions (ET, OU, NON … dite aussi algèbre de Boole, d'où le nom parfois employé de variables booléennes) ; pour cette raison, nous l'appellerons type *logique*. Il est merveilleusement adapté au calcul sur ordinateur, parce que c'est précisément le seul genre de nombre qu'un ordinateur reconnaisse, c'est comme qui dirait sa langue maternelle.

Récapitulons

Le tableau suivant résume les cinq types de variables qui ont été décrits ci-dessus, avec une abréviation et des paramètres, c'est-à-dire un système de contraintes qui encadre les valeurs possibles.

type	abrégé	description	paramètres
Calendaire	Cal	Date au format 12/02/2005	Implicites
Numérique	Num	Nombre réel	Minimum, maximum, précision[1]
Logique	Log	Deux valeurs antagonistes	Implicites[2]
Ordinal	Ord	Catégories ordonnées	Liste d'étiquettes[3]
Nominal	Nom	Catégories non-ordonnées	Liste d'étiquettes

Notes :

1) La précision est le nombre de chiffres significatifs autorisés *après* la virgule. Un nombre entier a donc une précision de *0* chiffre après la virgule, et – accrochez-vous – des nombres qui vont de 100 en 100 ont une précision de *-2*, puisqu'elle est alors de deux chiffres *avant*, et non *après* la virgule. La précision, ainsi que les extrema (minimum et maximum), sont des contraintes utiles au traitement informatisé : elles servent à la fois à encadrer le contrôle des erreurs en saisie et à optimiser la représentation interne des nombres, et donc la vitesse des calculs.

2) Les valeurs d'une logique sont implicitement Vrai et Faux.

3) Une étiquette est un texte arbitraire, choisi pour être compris par le lecteur des tableaux et résultats statistiques. Pour des raisons d'encombrement des tableaux, les étiquettes supportent, selon les logiciels, diverses contraintes ; dans Hector, être long de 10 caractères au maximum, espaces compris, tous caractères autorisés, sauf le caractère souligné '_'. C'est dire qu'avec 10 caractères on a intérêt à se doter d'un système d'abréviations performant. La liste des étiquettes associées à une variable nominale ou ordinale doit être définie à l'avance, car la saisie se fait par choix dans une liste.

Exercices

Il est d'une importance capitale d'être capable d'énoncer sans hésitation de type le variable correspondant à une information quelconque.

La page suivante vous propose plusieurs séries d'informations, dont vous devez essayer de déterminer le type s'il fallait les coder pour en faire une étude statistique.

Vous pouvez imprimer cette page, et, pour chaque information, cocher la case dont l'abréviation correspond au type de variable qui vous semble approprié.

Allons-y

	Cal	Num	Log	Ord	Nom
Le système religieux ou philosophique auquel vous adhérez					
La date de naissance de votre grand-père maternel					
La mention que vous avez obtenue au baccalauréat					
Votre intention de partir en voyage cet été					

Et encore

	Cal	Num	Log	Ord	Nom
Le numéro 59 sur les anciennes plaques d'immatriculation					
Le temps qui va s'écouler avant que vous ayez les réponses					
Le fait que vous approuviez ou pas la répression tabagique					
L'intérêt pour vous des statistiques (nul, bof, pas mal, super)					

On continue ?

	Cal	Num	Log	Ord	Nom
L'éventuel syndicat dont vous êtes membre					
La date exacte de votre premier rendez-vous amoureux					
Le degré de qualité du dernier Harry Potter					
Le temps qu'il vous faut pour boire une chope de bière glacée					

Si t'en re-veux, y'en re-n'a !

	Cal	Num	Log	Ord	Nom
La nationalité de votre grand-père paternel					
Les adverbes qu'on dit en effeuillant la marguerite					
Le fait que vous soyez célibataire ou non					
Le jour exact où vous avez prononcé vos premiers mots					

Un dernier pour la route

	Cal	Num	Log	Ord	Nom
Le nombre de copies que je peux corriger en une heure					
L'appréciation qui salue le goulash de Mémé					
Le jour de Noël					
La pointure des escarpins de la tante Emma					

Et pour finir

Tout le monde sait ce qu'une suite de nombres telle que 1460662119030 représente : c'est à l'évidence un numéro d'identification INSEE, ou numéro de sécurité sociale.

Indication : ça se segmente en 1 46 06 62 119 030 Mais saurez-vous identifier chaque segment et de quel *type de variable* il relève ?

Attention ! Tout ce qui s'écrit en chiffres n'est pas un nombre ! La paléo-informatique ne codait qu'en chiffres, ce qui en a embrouillé plus d'un(e).

Solutions et commentaires

Le système religieux est ici nominal : il n'y a pas d'ordre entre eux. Vous pouvez aussi n'en avoir aucun, ce qui peut se régler avec l'existence d'une étiquette 'aucun', ou avec une valeur absente.

La date de naissance est bien sûr calendaire.

La mention au bac est ordinale : très Bien vient après Bien.

L'intention de partir en voyage est une logique : il ne s'agit pas de quel voyage vous voulez faire, mais si vous voulez en faire un.

Le numéro sur les plaques est un ancien numéro d'ordre dans une liste alphabétique de départements. Mais de nouveaux départements sont apparus, hors de cet ordre, et certains (en Corse) ne sont plus repérés par un nombre. En fait ce n'est plus qu'une nominale, ici synonyme de « Nord », comme 33 est synonyme de « Gironde ».

Le temps qui va s'écouler est numérique, en jours, heures, minutes ou secondes, peu importe puisque ce sont multiples les uns des autres.

Le fait d'approuver ou pas est logique.

L'intérêt des statistiques, avec une série d'appréciations croissant vers le positif, est ordinale.

L'éventuel syndicat est nominal

La date exacte est calendaire

Le degré de qualité, comme toute évaluation, est ordinal

Le temps qu'il faut est numérique, en secondes ou en minutes, question de tempérament

Une nationalité est nominale

Les adverbes de la marguerite ne sont ordinaux que si on commence par « pas du tout », sinon il y a désordre, et c'est nominal

Le fait d'être célibataire ou non est une logique

Tout jour exact est une calendaire

Le nombre de copies est numérique

L'appréciation du goulash est ordinale

Le jour de Noël est un piège : de quelle année ?

La pointure des escarpins, ça dépend de l'origine nationale des chaussures. En France, un point de pointure vaut 2/3 de cm (une pompe en 42 fait 28 cm de long), c'est donc une numérique. Dans les pays anglo-saxons, il semble plutôt qu'on ait affaire à une ordinale, avec des tailles comme 5, 6, 7 dont les différences ne sont pas proportionnelles ; il s'agit alors d'une ordinale, même si l'existence de demi-tailles peut troubler.

Dans 1 46 06 62 119 030

1 est une logique, puisque 2 valeurs de genre seulement sont admises. A noter que par convention, on ne code pas le genre en logique, mais en nominale à deux valeurs, car une logique équivaudrait à « Est-un-homme, oui ou non ? » ou l'inverse, de quoi ne pas se faire d'ami(e)s dans les deux cas.

46 est un millésime cyclique, qu'on ne peut plus traiter comme une numérique, puisque 10 dénote 2010 alors que 1946 dénote 1946 … jusqu'en 2046. Alors, c'est quoi ? Une monstruosité, imprévue au départ, comme le bug (pipeau) de l'an 2000

06 est un nombre puisque c'est le numéro du mois dans l'année

62 est comme le 59 des plaques un ancien numéro d'ordre, maintenant nominal, qui dénote une commune

119 est un numéro d'ordre de commune dans une liste départementale qui ne se confond pas avec le code postal. Il est sans doute plus raisonnable de tenir le groupe 62119 comme une unique variable nominale.

030 est un nombre : le rang dans l'ordre des naissances de cette commune ce mois-là, ce qui fait de ce code entier celui du 30ème bébé, un garçon, né en juin 1946 à Béthune (Pas-de-Calais). *Ch'tot mi*. Comment font-ils avec les gosses qui naissent dans les avions ?

Trier une variable

C'est l'opération statistique fondamentale, qui consiste, devant un ensemble de choses qui ne sont pas toutes pareilles, à compter combien il y en a de chaque sorte ; par exemple, dans un tiroir, compter combien il y a de chaussettes de chaque couleur. En fait, il y a là deux opérations distinctes et successives : trier, c'est-à-dire rassembler ce qui est pareil et séparer ce qui n'est pas pareil, et compter, c'est-à-dire associer à chaque tas résultant du tri le nombre d'éléments qu'il contient. Pratiquement, parce qu'on fait presque toujours les deux opérations en série, on en est venu à parler de tri pour ce qui est en fait un tri-et-comptage.

Le tri, comme on le considère ici, concerne une seule variable à la fois, et non pas les relations entre plusieurs variables, ce qui sera l'affaire des croisements. C'est donc l'instrument principal de la statistique *descriptive*, celle qui montre mais ne suppose pas.

Du même tri, on peut la plupart du temps tirer plusieurs résultats, ou plusieurs présentations des mêmes résultats :

- une *table*, ou résultat tabulaire
- une *statistique locale*, ou de détail souvent logée dans une extension de la table
- une *statistique globale*, qui propose un ou des paramètres descriptifs généraux de la distribution étudiée
- un *graphe*, ou résultat graphique, qui représente la distribution comme un dessin

La section précédente a mis en évidence les différents types de variables : calendaire, numérique, logique, ordinal, nominal. Les résultats qu'on peut obtenir d'un tri dépendent du type de la variable qui est triée.

Tri d'une variable numérique

(score brut)

	effectifs	%/Total	% cumulés
0	3	0,60%	0,60%
1	23	4,56%	5,16%
2	44	8,73%	13,89%
3	77	15,28%	29,17%
4	89	17,66%	46,83%
5	89	17,66%	64,48%
6	78	15,48%	79,96%
7	50	9,92%	89,88%
8	34	6,75%	96,63%
9	13	2,58%	99,21%
10	4	0,79%	100,00%
Total	504	100.00%	

Ci-dessus, la *table*, enrichie de la statistique locale (les deux colonnes de droite).

La première colonne est celle des valeurs (les x_i), la seconde celle des effectifs (les n_i), c'est-à-dire du nombre de sujets (ou d'observations) relevant de la valeur de la même ligne *i*.

Ces deux premières colonnes constituent la table proprement dite, qui est toujours la liste des valeurs accompagnées des effectifs concernés. Les deux suivantes sont les *statistiques locales*.

La troisième colonne est celle du pourcentage que l'effectif n_i représente par rapport à l'effectif total (ligne du bas), souvent noté *N*. Si on aime les notations statistiques, on peut écrire :

$$pourcentage_i = 100 \times \frac{n_i}{N}$$

La quatrième colonne contient les pourcentages cumulés, c'est-à-dire, pour chaque ligne, la somme des pourcentages contenus dans elle-même et toutes les lignes précédentes, ou, si l'on préfère :

$$pourcentage\_cumulé_i = \sum_{j=1}^{j=i} pourcentage_i$$

Cette notation n'a rien d'effrayant en soi : elle décrit en fait une action, celle qui consiste à additionner tous les $pourcentages_j$, sachant que j parcourt tous les numéros de ligne de 1 à i.

Dans la table présentée ci-dessus, il y a une ligne pour chaque valeur différente ; si celles-ci étaient trop nombreuses au regard d'un réglage qu'on trouve dans les options (du genre : pas plus de quarante lignes dans une table numérique), le logiciel procéderait automatiquement à des regroupements en *classes*, c'est-à-dire en série de valeurs sur une même étendue. Avec par exemple une centaine de valeurs, on aurait sans doute, au lieu des valeurs, les classes [0 , [5 , [10 … désignant les classes [0 5[, [5 10[, etc.

Statistiques globales

```
Valeur modale : 4 (n=89)
```

La *valeur modale* est la valeur pour laquelle l'effectif est le plus grand, ou valeur la plus probable. Dans l'histogramme ci-dessous, c'est le sommet de la courbe. Si les valeurs sont regroupées en classes, on aura une *classe modale* plutôt qu'une valeur.

Parfois cette valeur n'est pas unique : il y a plusieurs valeurs offrant le même effectif maximum, voire des maxima proches. La distribution est alors dite plurimodale. C'est d'ailleurs le cas ici, puisque la valeur 5 a, par coïncidence, exactement le même effectif de 89 que la valeur 4

```
Médiane entre 4 et 5
```

La médiane possède plusieurs définitions concurrentes, et les différents ouvrages ne sont pas nécessairement en accord sur ce point. Pour ma part, je préfère celle-ci : la médiane est la coupure entre deux valeurs consécutives (ou le couple de ces deux valeurs consécutives), telle que le pourcentage cumulé de la première des deux valeurs soit la plus proche possible de 50%. Ici, elle est entre 4 et 5, parce que 46,83% est plus proche de 50% que ne l'est 64,48%. D'autres auteurs diraient ici que la médiane est à 4,5, voire à une valeur interpolée plus finement entre 4 et 5 en fonction des effectifs de part et d'autre.

```
Moyenne 4.74, écart-type 2.04
```

La moyenne est une des statistiques les plus (sinon les mieux) connues, puisqu'on en fait un usage immodéré dans les contextes scolaires. C'est la somme des valeurs, pondérées par les effectifs correspondants, divisée par l'effectif total :

$$m = \sum \frac{n_i x_i}{N}$$

on ne précise pas l'indice de variation du Σ, parce que comme il n'y a qu'un seul indice i dans la formule, c'est implicitement « pour toutes les valeurs de i ».

La moyenne peut aussi être notée m_x ou $\bar{x}$ (prononcé : x-barre).

La moyenne est un paramètre de tendance centrale, l'écart-type est un paramètre de dispersion : on le définit comme la racine carrée de la moyenne des carrés des écarts à la moyenne :

$$s = \sqrt{\frac{\sum n_i (x_i - m)^2}{N}}$$

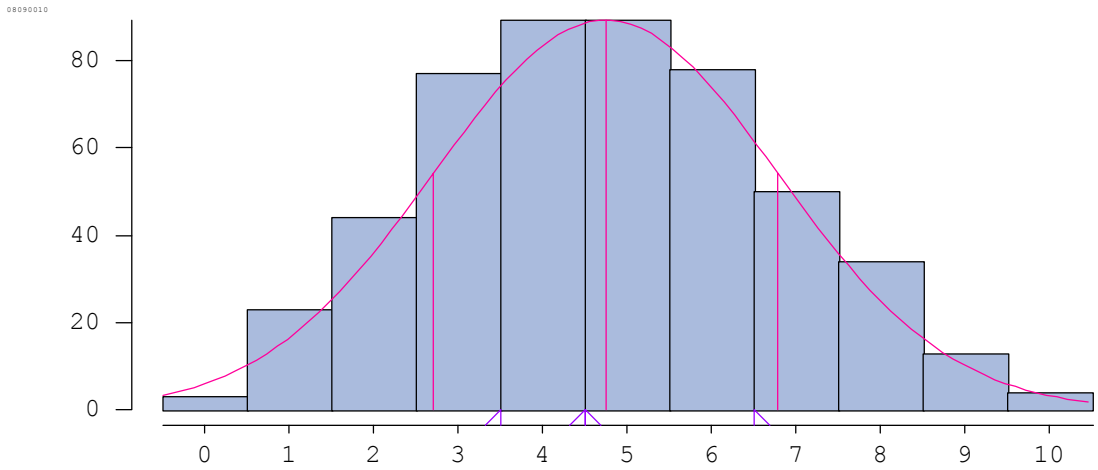
Cela se note *s*, (comme *standard deviation*), ou σ (sigma minuscule) si on est très snob.

L'élévation au carré des différences, avant l'addition, a pour objet d'empêcher que des différences positives et négatives se neutralisent mutuellement ; la racine carrée globale (le terme sous la racine est aussi appelé *variance*) pour revenir dans le même système d'unités que les valeurs.

Le couple de paramètres constitué par la moyenne et l'écart-type constitue un résumé suffisamment fidèle de la distribution, à condition que celle-ci soit approximativement normale, en gros : symétrique, avec un seul sommet arrondi au milieu, des extrémités effilées… Si ce n'est pas le cas (dissymétrie, bimodalité …), l'usage de la moyenne et de l'écart-type comme résumés peut conduire à de graves erreurs. Ces deux paramètres constituent la base des statistiques *paramétriques* ; quand on ne peut pas les utiliser, par exemple pour cause de non-normalité des distributions, on a recours aux statistiques non-paramétriques (voir ci-après).

Graphe

Le graphe usuellement associé à une variable numérique est *l'histogramme* :



Dans ce graphique, l'échelle horizontale (abcisses) correspond aux valeurs (les x_i). Si ces valeurs sont trop nombreuses, elles peuvent être regroupées en classes. L'échelle verticale (ordonnées) correspond aux effectifs pour chaque valeur ou classe de valeurs (les n_i). Pour chaque valeur ou classe de valeurs, l'histogramme comprend un rectangle dont la hauteur est proportionnelle à l'effectif concerné.

Dans la représentation graphique ci-dessus, une option a été utilisée, qui permet de superposer à l'histogramme la courbe idéale que dessinerait une distribution normale de même moyenne, de même écart-type et de même mode. Cette option provoque également l'affichage, sur l'axe horizontal, de repères correspondant à la médiane (coupure 50/50) et aux interquartiles (coupures 25/75 et 75/25). Il s'agit là d'enrichissements de la représentation graphique, et non de caractéristiques fondamentales de l'histogramme.

On ne présente pas ici le tri d'une variable calendaire, qui est une simple variante de variable numérique, avec une convention particulière de notation des valeurs (modèle jj/mm/aaaa).

Tri d'une variable ordinale

Une variable ordinale a pour valeurs des « étiquettes », textes arbitraires mais supposés faire sens pour le lecteur, entre lesquelles existe un ordre significatif.

La *table* est donc presque la même que pour une numérique (les valeurs sont des textes et non plus des nombres ou des classes d'intervalles de nombres), et la *statistique locale* comporte des pourcentages cumulés de haut en bas, puisque l'ordre fournit un haut et un bas.

(groupe de résultats)

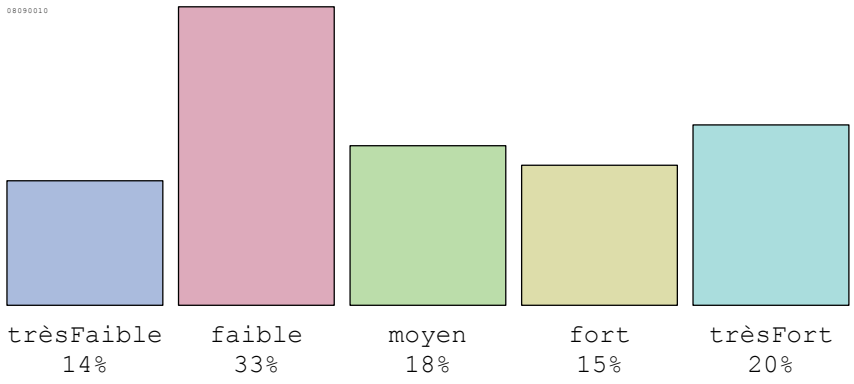
	effectifs	%/Total	% cumulés
trèsFaible	70	13,89%	13,89%
faible	166	32,94%	46,83%
moyen	89	17,66%	64,48%
fort	78	15,48%	79,96%
trèsFort	101	20,04%	100,00%
Total	504	100.00%	

Valeur modale : faible (n=166)

Médiane entre faible et moyen

La *statistique globale* admet aussi un mode et une médiane, toujours à cause de l'ordre, mais pas de moyenne (ni d'écart-type) parce que les intervalles entre valeurs ne sont pas comparables entre eux : l'écart entre moyen et fort n'est pas comparable à l'écart entre fort et très fort, et il n'est pas possible d'en compenser l'un par l'autre, et la notion de moyenne, ou des valeurs fortes compensent des valeurs faibles, n'a pas de signification.

Le *graphe* est ici un diagramme en barres verticales, dont la hauteur est proportionnelle à l'effectif concerné, l'ordre des valeurs étant conservé de gauche à droite.



Tri d'une variable logique

Les valeurs d'une variable logique sont exclusivement Faux et Vrai, mais pour distinguer les valeurs de deux variables logiques, on a remplacé ici Faux et Vrai par des étiquettes de synthèse, formées d'une sorte de résumé de l'intitulé de la variable (les premières lettres des mots qui le composent), suivi du signe – ou + pour Faux et Vrai.

La *table* a une allure maintenant classique, et les statistiques locales se limitent aux pourcentages, puisqu'un cumul sur deux classes seulement n'offre aucun intérêt.

(ortho)

	effectifs	%/Total
Ortho-	311	61,71%
Ortho+	193	38,29%

Total	504	100.00%

Intervalle de confiance à .05 : [36,23% 40,36%]

La *statistique globale* est l'intervalle de confiance, qui repose sur l'idée suivante : le pourcentage de Vrais observé, ici 39.29% ou encore f = .39 si on le note en fréquence, peut être considéré comme une valeur d'échantillon d'une fréquence F théorique dans une population parente de grande taille.

Cette fréquence théorique est une variable aléatoire de moyenne F et d'écart-type :

$$s = \sqrt{\frac{f(1-f)}{N}}$$

Cela permet de déterminer une sorte de fourchette entre lesquelles peut naviguer la valeur observée : F-zs et F+zs, où z est une valeur normale réduite correspondant au seuil de probabilité choisi. On utilise ici le seuil conventionnel en sciences humaines P =.05, pour lequel z = 1.96 (lecture des tables de la Loi Normale).

Le *graphe* est une simple barre horizontale découpée proportionnellement aux fréquences de Vrai et de Faux.

On peut évidemment discuter de l'intérêt informatif d'un tel graphe. Disons qu'il est là parce qu'il fallait qu'il y eût un graphe dans tous les cas …

Tri d'une variable nominale

Une variable nominale a pour valeurs des étiquettes, comme la variable ordinale, mais aucun ordre significatif n'existe entre ces valeurs.

La *table* est classique, et la *statistique locale* se réduit aux pourcentages, car aucun cumul n'aurait de signification.

(origine des données)

	effectifs	%/Total
eLearn05	140	27,78%
eLearn06	110	21,83%
LicSE0304	61	12,10%
Ortho06	96	19,05%
Ortho07	97	19,25%
Total	504	100.00%

Efficacité entropique : 98,0%

La *statistique globale* est l'efficacité entropique, issue des théories de l'information. Elle fonctionne comme une mesure d'homogénéité de la distribution, c'est-à-dire de tendance des différentes lignes à contenir des effectifs proches. L'idée que se fait cette mesure d'une distribution efficace est une distribution où toutes les valeurs ont exactement le même effectif.

C'est une certaine conception de l'efficacité et ça ne prétend pas être autre chose.

L'efficacité entropique se calcul comme le rapport entre H, entropie observée dans le tableau,et $\max H$, l'entropie maximale que peut produire un tableau du même nombre k de valeurs différentes. Elle vaut donc 1 quand l'entropie est égale au maximum possible :

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i$$

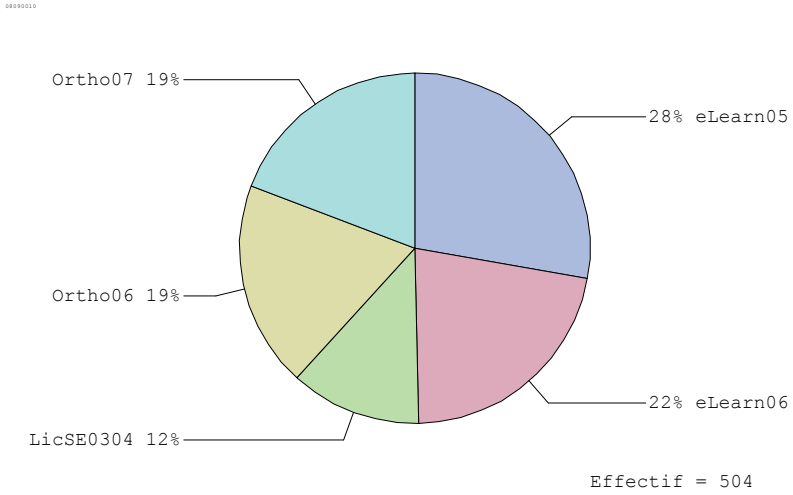
où p_i, fréquence d'une case, est le rapport n_i/N, et où les logarithmes utilisés sont en base 2

et

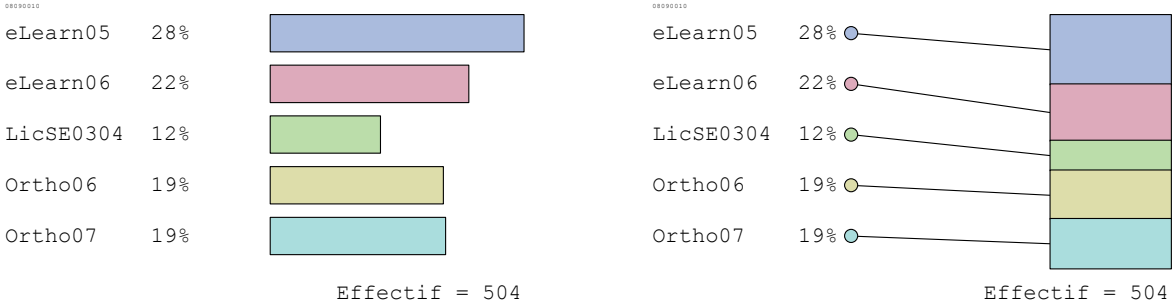
$$\max H = \log_2 k$$

où k est le nombre de valeurs distinctes.

Plusieurs types de *graphe* sont envisageables pour le tri d'une nominale. Ceci est le diagramme en secteurs, ou « camembert ». Chaque secteur a un angle proportionnel à l'effectif d'une valeur. La circularité met en évidence l'absence d'ordre.



Ci-dessous, le diagramme en barres et le diagramme en couches



Dans un cas, c'est la longueur des barres qui est proportionnelle aux effectifs, dans l'autre, c'est l'épaisseur des couches. Le choix d'une ou l'autre méthode dépend de ce qu'on souhaite mettre en évidence et illustrer particulièrement.

Croiser deux variables : 3 cas de figure et pas un de plus

Contexte : deux types fondamentaux, numérique et nominal

Dans les pages précédentes, on a distingué cinq types de variables : calendaire, numérique, logique, ordinal, nominal.

Pour distinguer l'essentiel de l'accessoire, on ne considère plus, à partir de ce point, que deux types fondamentaux et antagonistes :

- le type numérique code des mesures ou des dénombrements, et s'exprime par des nombres, autorisant donc une grande diversité d'opérations mathématiques (notamment le calcul de la moyenne). Exemples : la taille, le poids, la note en philosophie …
- le type nominal, qui code l'appartenance à une catégorie (groupe ou classe) par un texte, une étiquette arbitraire. Exemples : le sexe, la couleur des cheveux, la nationalité …

Le type calendaire est laissé de côté car il s'agit d'une simple variante d'une numérique, caractérisée par le format de date jj/mm/aaaa. Le type ordinal est un bâtard du nominal,

auquel il emprunte le codage par étiquettes, et du numérique auquel il emprunte la relation d'ordre ; il mériterait à lui seul un autre développement. Le type logique a toutes les vertus du nombre et de l'étiquette, et exploite l'une ou l'autre selon les situations ; il n'est donc pas utile de le considérer ici comme un type fondamental.

Croiser deux variables, qu'est-ce que c'est, et pour quoi faire ?

On considère ici comme suffisamment connus (on y reviendra si nécessaire) les éléments relatifs au tri d'une seule variable à la fois.

Nous appelons *croisement* le tri des sujets selon deux variables simultanément.

Le nom de croisement vient du fait qu'une manière classique d'en représenter le résultat est le *tableau croisé*, c'est-à-dire un tableau où se *croisent* les traits horizontaux qui séparent les valeurs d'une des variables et les traits verticaux qui séparent les valeurs de l'autre.

(classe d'âge) x (classe de résultats)

N	insuff	médiocre	satisf	supérieur	S/LIGNE:
vétérans	42	46	60	56	204
mûrs	63	55	48	46	212
jeunes	58	41	40	32	171
benjamins	38	49	55	71	213
S/COLONNE:	201	191	203	205	800

L'opération de tri consiste à considérer successivement chacun des sujets, ou observations, et à décider, selon ses caractéristiques, dans quelle case on va le compter. Par exemple, on suppose que le premier sujet appartient, du point de vue de la *classe d'âge*, à la catégorie *jeunes* ; mais qu'il appartient aussi, du point de vue de la *classe de résultats*, à la catégorie *médiocre* ; aussi doit-il être compté dans la case qui est à l'intersection de la ligne *jeunes* et de la colonne[113] *médiocre*, celle qui contient 41 sujets à la fin du tri croisé.

Le tableau ci-dessus est le résultat final de l'opération : on peut s'imaginer le logiciel comme un employé du tri postal qui répartit à toute vitesse les 800 sujets dans les cases, et à la fin dit combien il y en a dans chaque case : 42 relèvent de la *classe d'âge* étiquetée *vétérans* ET de la *classe de résultats* étiquetée *insuff*, 29 relèvent de la *classe d'âge* étiquetée *vétérans* ET de la *classe de résultats* étiquetée *médiocre,* etc.

Le tableau croisé proprement dit n'est constitué que des 16 cases (4 lignes x 4 colonnes) du centre non grisé. La première ligne et la première colonne contiennent les valeurs des variables, ici des étiquettes ; la ligne du bas et la colonne de droite sont des marges : celle-ci contient les effectifs totaux par lignes, celle-là les effectifs totaux par colonnes ; la dernière case en bas à droite, qui est leur intersection, contient évidemment l'effectif total général.

En résumé, croiser deux variables c'est donc ventiler les sujets (ou observations) simultanément selon deux critères, et compter les effectifs correspondants à chaque combinaison de critères.

Pourquoi se livrer à cette activité ? *Parce qu'on a quelque chose derrière la tête*.

Si l'on croise une variable comme la *classe d'âge* avec une variable comme la *classe de résultats*, c'est parce qu'on subodore que, selon la classe d'âge, on ne se répartit pas de la même façon dans les classes de résultats, et en tous cas on aimerait le vérifier. Si on y parvient, on aura prouvé l'existence, en tous cas dans cette population, d'une relation entre la variable *classe d'âge* et la variable *classe de résultats*.

Tout est là, et c'est toute la *statistique inductive* : en partant d'irrégularités de répartition entre différents cas de figure, je vais induire l'existence d'une relation entre

[113] Les lignes sont horizontales : penser à la ligne d'horizon ; les colonnes sont verticales : penser aux colonnes du Parthénon.

deux phénomènes. Contrairement à la déduction, où on applique aux faits des lois connues, dans l'induction, on part des faits pour essayer d'en dégager les lois.

Le croisement selon les types

Nous avons réduit le nombre de types fondamentaux à deux : numérique et nominal. A priori, cela engendre quatre types de croisements de deux variables : nominal × nominal, nominal × numérique, numérique × nominal, numérique × numérique. En réalité, cela n'en fait que trois, parce que le second et le troisième sont équivalents[114].

Dans tous les cas, nous continuons, comme avec le tri, de considérer quatre types de résultats[115] :

- un résultat tabulaire, ou *table* : c'est le tableau croisé lui-même ou quelque chose qui en est dérivé
- une *statistique locale* : ce sont des éléments statistique de détail, qui facilitent l'interprétation des résultats. La statistique locale est parfois nichée dans le résultat tabulaire, parfois indépendante
- une *statistique globale* : c'est un test statistique qui fournit diverses informations sur la vraisemblance d'une relation entre les deux variables
- un résultat graphique, ou *graphe* : c'est une représentation graphique de la table.

Ces quatre types de résultats, qu'ont peu inclure ou exclure de ce qui est affiché, jouent dans le traitement des données un rôle fondamental.

L'idée de base est que le hasard[116] est un farceur.

Si j'observe la descente des passagers du ferry-boat à Calais, avant que le *shuttle* ne l'ait ruiné, je peux repérer, parmi les représentantes de la moitié agréable du genre humain, qui est anglaise et qui ne l'est pas (à la voix), qui est rousse et qui ne l'est pas (à la tignasse).

Ce jour-là débarquent 190 passagères (c'est peu, mais c'est l'hiver, et pour traverser le *Channel* dans ces conditions, il faut une sérieuse motivation). Parmi elles, 40 sont des anglaises rousses, 50 des anglaises blondes, brunes, ou n'importe quelle autre teinte, bref, pas rousses, 20 sont rousses, de nationalités diverses mais pas anglaises, 80 ne sont ni rousses ni anglaises. Mes 190 observations se résument dans un tableau croisé de ce genre :

	rousses	Pas rousses
anglaises	40	50
autres	20	80

Ca, c'est la *table brute*. Rien que des effectifs.

Je suis évidemment tenté de dire qu'il y a vraiment beaucoup d'anglaises qui sont rousses. Pourtant, me dira-t-on, il y a plus d'anglaises pas rousses que d'anglaises rousses. Oui, mais la majorité des rousses sont anglaises. Ah, j'ai parlé de majorité, c'est que je suis sur le point de raisonner sur des *proportions* plutôt que sur des *effectifs* absolus.

---

[114] Quand on croise une numérique et une nominale, la nominale se met forcément en premier : on verra pourquoi.

[115] Il faut y voir l'esprit de système chez l'auteur de ce texte, qui est aussi le développeur du logiciel Hector utilisé pour les exemples. Dans certaines circonstances, tel ou tel résultat peut présenter moins d'intérêt fondamental que tel autre.

[116] Selon le rapport qu'on entretien avec la nature et les croyances qu'on interpose entre soi-même et l'absurdité de la destinée humaine, on peut se représenter le hasard comme un lutin sympathique, un gremlin vicieux, une calamité météorologique, un cheïtan maléfique, un faiseur de coïncidences ou un ennemi de la planification rationnelle … et sans doute bien d'autres rôles encore

Pour raisonner sur des proportions, rien de tel que des pourcentages :

	rousses		Pas rousses	
anglaises	40	*44,44%*	50	*56,56%*
autres	20	*20,00%*	80	*80,00%*

Mais 44,44% de quoi ? Ce tableau est incomplet : où sont les 100% ?

	rousses		Pas rousses		Ensemble	
anglaises	40	*44,44%*	50	*56,56%*	90	*100,00%*
autres	20	*20,00%*	80	*80,00%*	100	*100,00%*
Ensemble	60	*31,58%*	130	*68,42%*	190	*100,00%*

Les voilà : au bout de chaque ligne. On les appelle pour cette raison des *pourcentages-lignes*. Ils permettent de répondre à la question : parmi les *anglaises* d'une part, et les *autres* d'autre part, quelle proportion trouve-t-on de *rousses* et de *pas rousses* ? Et de répondre qu'en proportion, il y a plus de deux fois plus de *rousses* parmi les *anglaises* (44,44%) que parmi les *autres* (20% seulement).

On notera que poser la question comme ça, c'est placer la nationalité plutôt du côté des *causes*, et la couleur de cheveux du côté des *effets*. On dirait aussi que la variable nationalité est placée en *variable indépendante*, et la variable couleur des cheveux en *variable dépendante*.

Pouvait-on poser une autre question ? Certainement :

	rousses		Pas rousses		Ensemble	
anglaises	40	*66,67%*	50	*38,46%*	90	*47,37%*
autres	20	*33,33%*	80	*61,54%*	100	*52,63%*
Ensemble	60	*100,00%*	130	*100,00%*	190	*100,00%*

Cette fois ce sont des *pourcentages-colonnes* : le 100% est au pied des colonnes. La question qu'on peut se poser est : parmi les *rousses* d'une part et les *pas rousses* d'autre part, quelle est la proportion d'*anglaises* ou d'*autres*. Cette fois, c'est la couleur des cheveux qui est renvoyée du côté des causes, variable indépendante, et la nationalité du côté des effets, variable dépendante.

Quelle question est la plus légitime ? Ca dépend de ce qu'on veut démontrer. La première paraît un tout petit peu plus logique, mais sans plus, après tout ça n'est qu'un exercice. Dans certains cas, la relation cause-effet semble évidente, dans d'autre moins.

En tout cas, s'il n'y a pas de bonne question, il existe une *manière conventionnelle* d'organiser ce genre de tableau : on place la *variable indépendante en lignes*, et la *variable dépendante en colonnes*, et on utilise les pourcentages-lignes. L'intérêt de ce type de convention est qu'il diminue les efforts de lecture, comme celle qui dit que l'écriture des langues occidentales se lit de gauche à droite[117]. On a donc intérêt à placer comme première variable, qui fournira les lignes, celle que l'on considère plutôt comme la variable indépendante. On peut faire le contraire, quand par exemple le nombre de lignes est raisonnable mais le nombre de colonnes excessif : on échange alors les deux variables et on utilise les pourcentages-colonnes, mais *on doit en avertir le lecteur*.

[117] Le fait qu'il s'agisse d'une convention est attesté par l'existence, dans certaines variantes précoces du grec, d'une écriture dite *boustrophédon*, par imitation du laboureur qui fait faire demi-tour aux bœufs attelés à la charrue au bout du sillon, où au bout de la ligne on entamait la suivante dans l'autre sens, plutôt que de revenir à la marge gauche

Dans cet exemple, les pourcentages-lignes jouent le rôle d'une *statistique locale*, dont le rôle est de faciliter l'interprétation détaillée des résultats.

Puis-je enfin proclamer mes résultats, annoncer que la proportion de rousses parmi les anglaises ce jour-là est étrangement élevée, et attirer l'attention de la Police de l'Air et des Frontières sur une possible infiltration irlandaise ?

Non pas, il me reste un souci. Et si le hasard me faisait une blague ?

Je sais bien que quand je joue à pile ou face, la probabilité est de 50/50. Pourtant, sur une série de dix lancers, bien d'autres résultats que 5/5 peuvent apparaître, et même une série de 10 piles ou de 10 faces est rare, mais pas strictement impossible. Si elle survient, est-ce que ça veut dire que la pièce est truquée, ou que mon adversaire triche ? Pas nécessairement : même si l'on sait ce que *devrait* être le résultat (sa loi de probabilité), la réalité d'un échantillon d'observation peut s'en écarter notablement, et d'autant plus facilement qu'il est de petite taille.

Alors, si je me tourne vers la statistique, va-t-elle enfin m'apporter des *certitudes* ?

Hélas non ! Autant en faire son deuil. La statistique, inductive en l'espèce, n'apporte *jamais* de certitude.

A quoi sert-elle alors, cette feignasse ?

La statistique apporte quelque chose de presque aussi utile qu'une certitude : *la mesure du risque d'erreur que je prends quand j'affirme quelque chose*.

Le raisonnement est le suivant : nos prédécesseurs mathématiciens ont étudié les lois de probabilité dans beaucoup de cas de figure, et leurs travaux permettent de mesurer à quel point ce qu'on observe est rare, si c'est seulement dû au hasard.

Par exemple, de tel phénomène, les lois de probabilité me diront qu'il peut se produire aléatoirement dans moins d'un cas sur 100 : on parlera alors d'un seuil de probabilité[118] P = .01.

Si j'ai choisi de régler la sensibilité de ma *machine à méfiance* à ce seuil, je dirai qu'à .01 je n'y crois pas, qu'il faut chercher d'autres explications que le hasard à ce qu'on constate.

Ce faisant, comme dans la réalité ça peut quand même arriver aléatoirement dans 1% des cas, je prends et j'assume un risque de me tromper une fois sur 100. Ca paraît beaucoup, mais c'est ça ou ne jamais rien savoir, et ne jamais rien dire. C'est comme trouver une grosse crotte de chien exactement sur le seuil de sa porte tous les matins alors qu'on n'a pas de chien, et admettre que c'est sans doute dû au hasard.

Dans le cas de notre grande enquête sur les éphélides[119], comment ça fonctionne ?

On raisonne comme ceci : s'il n'y avait pas de lien particulier entre la couleur des cheveux et la nationalité des passagères du ferry-boat, alors on trouverait des rousses et des non rousses en mêmes proportions dans toutes les nationalités, et au lieu d'avoir 44,44% de rousses parmi les anglaises, on en aurait 31,58%, comme dans l'ensemble de l'échantillon :

	rousses		Pas rousses		Ensemble	
anglaises	40	*44,44%*	50	*56,56%*	90	*100,00%*
autres	20	*20,00%*	80	*80,00%*	100	*100,00%*
Ensemble	60	*31,58%*	130	*68,42%*	190	*100,00%*

Le tableau ressemblerait alors à ceci :

[118] .01 ou 1 sur 100 ou 1% ou 0,01 : tout ça c'est la même chose.

[119] Mot savant et très joli pour désigner les taches de rousseur.

	rousses		Pas rousses		Ensemble	
anglaises	28,42	*31,58%*	62,58	*68,42%*	90	*100,00%*
autres	31,58	*31,58%*	68,42	*68,42%*	100	*100,00%*
Ensemble	60	*31,58%*	130	*68,42%*	190	*100,00%*

Mais ce n'est pas possible d'avoir 28,42 rousses, on ne peut pas les couper en morceaux ! En effet : ce ne sont plus des *effectifs observés* comme dans le vrai tableau, mais des *effectifs théoriques*, des valeurs imaginaires qui correspondent à l'hypothèse où les deux variables sont *indépendantes*.

On va maintenant calculer une grandeur statistique qui sera d'autant plus grande que les valeurs du tableau observé s'écartent de celles du tableau théorique. Pour cela, dans chacune des cases du tableau, on fait la différence de l'observé et du théorique, on élève cette différence au carré pour que les différences positives et négatives ne se neutralisent pas mutuellement, on divise par l'effectif théorique pour relativiser, et ça donne la contribution de la case. On additionne enfin la contribution de toutes les cases, et on appelle ça le χ^2 (prononcer Khi-deux).

Dans ce cas précis, le calcul donne 15,27. C'est une valeur très élevée. Comment en déduire le seuil de probabilité, et donc le risque d'erreur ? On utilise une table[120], la table des χ^2, où est gravé ce qu'on sait du phénomène. Cette table possède plusieurs lignes, selon le nombre de degrés de liberté du tableau qu'on étudie. Pour le connaître, on enlève 1 au nombre de lignes, 1 aussi au nombre de colonnes, et on multiplie. Donc, ici, 1 degré de liberté. La table a plusieurs colonnes :

	.10	.05	.01
1	2,71	3,84	6,64
2	4,61	5,99	9,21

En tête de chaque colonne, un seuil de probabilité : .10 , .05, .01

Dans les colonnes, pour chaque ligne, des valeurs-seuil.

On lit le tableau à la ligne 1, puisqu'on a 1 degré de liberté. On aborde le tableau par la gauche, et on se déplace vers la droite tant que le χ^2 qu'on a calculé est supérieur ou égal au χ^2 seuil lu dans la table.

Dans notre cas, ça nous amène à sortir de la table par la droite, et à conclure que oui, au seuil P = .01, on peut rejeter l'Hypothèse nulle (le hasard serait responsable), et constater l'existence d'une relation entre la nationalité et la couleur des cheveux. La table aurait-elle pu avoir d'avantage de colonnes ? Oui, on aurait pu aller vers des seuils de probabilité encore plus fins, comme .005 , .001 … En Sciences Humaines, on se contente usuellement de ceux-là, mais en pharmacologie ce serait sûrement différent : tout dépend de la nature du risque[121].

Que se serait-il passé si le χ^2 avait été plus faible ? A 4,50 par exemple il aurait été supérieur au 3,84 de .05 , mais inférieur au 6,14 de .01. On aurait donc conclu seulement au seuil P = .05, et donc a l'existence d'une relation significative, mais pas *très* significative comme avec .01. Avec un χ^2 de 3,20, on aurait conclu à une relation *peu*

[120] On trouve toutes sortes de tables de ce genre à la fin des bouquins de statistiques. Les utilisateurs d'un logiciel n'en ont pas besoin, puisque celui-ci fait tout le boulot, mais il est bon, une fois pour toutes, de savoir de quoi il s'agit.

[121] Le plus gros risque que nous prenions ici est d'écrire des bêtises. D'autres que nous l'assument sans frémir.

significative, et à moins de 2,71 on n'aurait pas rejeté l'Hypothèse nulle, parce que plus de 10% d'erreur c'est quand même un trop grand risque, et dans ce cas on ne conclut rien du tout (χ^2 non significatif).

Le χ^2, avec le seuil de probabilité correspondant, nous sert ici de *statistique globale*.

Son rôle est de nous informer sur l'existence d'une relation significative, et donc sur notre droit de la commenter.

En d'autres termes, et pour essayer d'être tout à fait clair : si la statistique globale n'est pas significative, *on n'a pas le droit* scientifique de commenter le tableau, parce qu'il n'y a rien de spécial à commenter.

La démarche correcte est donc la suivante :

- On considère d'abord la *statistique globale*. Si elle n'est pas significative, on le dit et on passe à autre chose[122].
- Si la *statistique globale* est significative,
  - on examine de plus près le *résultat tabulaire*, éventuellement accompagné des *statistiques locales*,
  - et on s'en sert pour commenter,
  - en illustrant éventuellement à l'aide du *graphe*, ou résultat graphique, qui apporte les mêmes informations que la table et les statistiques locales, mais sous une forme résumée, parfois plus compréhensible.

Cette façon de procéder va se retrouver dans les différents cas de figure générés par les différents types de variable. On notera que plus le seuil de probabilité est fin (.01), plus le commentaire est assuré. A .05 on reste prudent et on essaye de confirmer par d'autres tableaux, à .10 on signale une tendance et on évite surtout tout raisonnement en chaîne[123].

Deux variables nominales

Considérons deux variables, issues d'un corpus des résultats de licence en sciences de l'Education en 1999.

La variable *filière d'accès* décrit comment les individus sont arrivés en licence : elle a pour valeurs :

- formGéné : diplôme de formation générale (DEUG ou autre licence)
- formProf : diplôme de formation professionnelle (BTS, DUT, diplômes de la formation d'adultes, de la santé et du travail social)
- expéProf : validation d'acquis de l'expérience professionnelle

La variable *classe d'âge* répartit les sujets en quatre catégories d'effectifs à peu près égaux, selon l'âge[124] :

- vétérans : 34 ans et plus
- mûrs : 27 à 33 ans
- jeunes : 24 à 26 ans
- benjamins : 20 à 23 ans

---

[122] Autrement dit, la statistique globale est l'arbitre, le juge de paix. Si elle est non-significative, le message est « Circulez, rien à voir ». Si on insiste quand même : « oui mais regardez, il y a tout de même une légère tendance … », on vient de sortir de la route de la démarche scientifique, et il faut s'apprêter à en subir les conséquences, notamment à l'examen ou en soutenance

[123] La valeur globale d'un raisonnement enchaîne ne saurait être supérieure à celle de son maillon le plus faible.

[124] Cette variable est en fait de type ordinal, mais cette propriété n'est pas exploitée ici : elle est traitée comme une nominale.

La *table brute* a l'allure suivante :

(filière d'accès) x (classe d'âge)

N	vétérans	mûrs	jeunes	benjamins	S/LIGNE :
formGéné	52	92	122	174	440
formProf	38	52	40	14	144
expéProf	75	52	13	4	144
S/COLONNE:	165	196	175	192	728

Poser le croisement dans cet ordre des variables, c'est se préparer à tester la question suivante : « est-ce que les différentes filières d'accès ont des affinités particulières avec différentes classes d'âge ? »

La table brute ne permet pas de répondre directement à cette question : il faut considérer d'abord la *statistique globale* :

Khi2 = 184,96 pour 6 d.d.l. , s. à .01

Le χ^2, calculé ici avec 6 degrés de liberté ((3 lignes -1) × (4 colonnes – 1)), est suffisamment élevé pour être significatif au seuil P = .01.

On peut donc, à ce seuil, rejeter l'Hypothèse nulle et admettre l'existence d'une relation très significative entre la filière d'accès et la classe d'âge.

Un χ^2 très significatif : on a donc le feu vert pour poursuivre.

Reste à savoir maintenant en quoi consiste cette relation très significative[125]. Pour cela, on examine les statistiques locales, qui, dans le cas nominale × nominale , consistent en l'ajout dans la table des pourcentages-lignes et des signes des associations locales.

(filière d'accès) x (classe d'âge)

N %L +	vétérans		mûrs		jeunes		benjamins		S/LIGNE :	
formGéné	52	12% ---	92	21% ---	122	28% +++	174	40% +++	440	100%
formProf	38	26%	52	36% +++	40	28%	14	10% ---	144	100%
expéProf	75	52% +++	52	36% +++	13	9% ---	4	3% ---	144	100%
S/COLONNE:	165	23%	196	27%	175	24%	192	26%	728	100%

Les signes des associations locales[126] sont des indicateurs de la force de l'affinité entre une ligne et une colonne, ou encore de la contribution de la case correspondante au χ^2 :

- +++ veut dire une association positive très forte : il y a beaucoup plus d'observations dans cette case qu'il n'y en aurait si les variables étaient indépendantes. ++ et + dénoteraient des associations positives plus faibles.
- à l'inverse, --- veut dire une association négative très forte : il y a beaucoup moins d'observations dans cette case qu'il n'y en aurait si les variables étaient indépendantes. -- et - , la même chose en moins fort.
- une absence de signe indique que l'effectif de la case est à peu près conforme au modèle de l'indépendance.

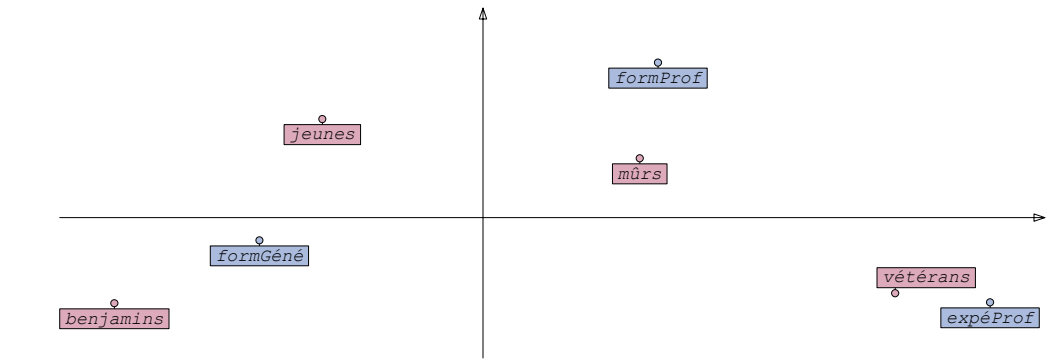
[125] Cette statistique a été calculée avec une version d'Hector qui n'examinait pas de seuil de significativité plus fin que .01. Recalculée avec une version plus récente qui cherche plus loin, on atteint ici le seuil P<.0000, ce qui veut dire qu'il n'y a pas de décimale significative avant la cinquième, autrement dit qu'on est aussi près que possible de la certitude.

[126] Une définition statistique sérieuse et détaillée du signe des associations locales est donnée dans l'Annexe statistique aux documentations du logiciel Hector. C'est trop technique pour qu'on s'y attarde ici. L'important est de s'en faire une représentation efficace.

Le tableau ci-dessus se caractérise donc par

- la forte association des formations générales aux deux classes d'âge les plus jeunes
- le lien entre les formations professionnelles et la classe d'âge des mûrs
- la correspondance entre l'expérience professionnelle et les deux classes les plus âgées.

Le choix du *graphe*, ou représentation graphique illustrative, est ensuite affaire de contexte et de goût :



93,06% de l'inertie sur l'axe 1 horizontal
6,94% de l'inertie sur l'axe 2 vertical

L'analyse factorielle de correspondances simple[127], assez spectaculaire, repose sur des éléments mathématiques qui sont au-delà des perspectives de ce texte d'initiation. Signalons simplement qu'il s'agit d'une représentation dans un espace à deux dimensions des affinités entre lignes et colonnes, qui sont traduites par des proximités dans le plan (une ligne et une colonne dont les étiquettes sont proches ont souvent +++ dans leur case commune dans le tableau).

La lecture de cette représentation insiste sur les jeux d'opposition portés par les axes, qui sont indépendants : d'abord l'axe horizontal, qui porte ici une opposition primaire entre jeunes et vieux, et la répartition des modes d'accès sur cette dimension ; le second axe, vertical, exprime la singularité de l'accès par la formation professionnelle.

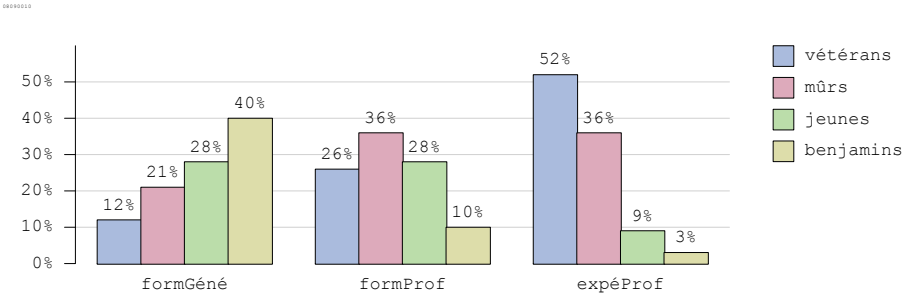
Les taux d'inertie répartis sur les axes disent la part d'information portée par chaque axe. Ici, le premier axe porte pratiquement tout : le fait important est donc bien le caractère générationnel des modes d'accès. La somme des deux taux est l'ensemble de l'information contenue dans le plan formé par les deux premiers axes. Si cette somme est inférieure à 100% , le reliquat correspond à des dimensions ultérieures du nuage de points, qui ne sont pas représentées dans le plan des deux premières dimensions. Si ce reliquat est important (de l'ordre de 50% par exemple), la portée illustrative de l'Analyse factorielle doit être fortement relativisée[128].

Attention ! L'analyse factorielle ne doit être commentée que sous couvert d'un $\div^2$ significatif : cette technique graphique est en effet une véritable machine à mettre en relief les différences, et elle le fait même si celles-ci sont extrêmement ténues et pas du tout significatives. En résumé, un joli graphique n'est pas une preuve !

---

[127] Plus précisément : analyse factorielle de correspondances illustrative d'un tableau de contingence.

[128] Ici le reliquat est nul, mais il ne pouvait en être autrement. Pour des raisons mathématiques, le nombre de dimensions distinctes du nuage de points représentant les lignes et le colonnes ne saurait être supérieur au plus petit nombre de degrés de liberté pris sur les lignes ou les colonnes. Or le nombre de lignes est de 3, soit deux degrés de liberté. Deux dimensions suffisent donc a exprimer toute l'information du tableau, et il n'y a pas de reliquat invisible.



On préférera souvent la représentation ci-dessus (barres et pourcentages), variante optionnelle dans le cas nominale × nominale . Le graphique est plus aisé à lire, et les associations positives fortes s'y manifestent comme les barres les plus hautes : l'allure de chaque ligne s'y lit comme un *profil*. Là aussi, bien sûr, le graphique sans statistique globale est inacceptable.

Dans un autre croisement du même genre, la table avec statistique locale, suivie de la statistique globale et du graphe, présente l'allure suivante :

(filière d'accès) x (classe de résultats)

N %L +	insuff	médiocre	satisf	supérieur	S/LIGNE :
formGéné	83 21%	85 22%	110 28%	110 28%	388 100%
formProf	22 17%	35 27%	33 25%	42 32%	132 100%
expéProf	25 20%	35 28%	26 21%	39 31%	125 100%
S/COLONNE:	130 20%	155 24%	169 26%	191 30%	645 100%

Khi2 = 5,66 pour 6 d.d.l. n.s.

Ici le χ^2 n'est pas significatif, et il n'y a pas lieu de commenter les résultats tabulaires. Aucune association locale ne se manifeste, et le graphe n'a pas non plus de raison d'être. Mais l'Analyse Factorielle, en bon petit soldat, est capable de le produire quand même. Introduire ce graphe dans un texte scientifique et pis encore le commenter serait une faute grave.

Ce qu'il convient d'écrire, c'est simplement : « le χ^2 non significatif ne permet pas de conclure à l'existence d'une relation entre les deux variables ». *Et puis c'est tout !*

Deux variables numériques

Considérons deux variables numériques, issues du même corpus que précédemment. A_NOTE et B_NOTE sont des notes obtenues dans des disciplines différentes, dont l'identité précise est sans intérêt ici.

(A_NOTE) x (B_NOTE)

N	0,0	1,0	2,0	3,0	4,0	5,0	…
0,0							
2,0	1		1				
3,0							
4,0				1			
4,5							
5,0						2	
6,0		1			3	3	
…							

Le tableau brut est très grand, comme souvent avec les variables numériques. On n'en a fait figurer ci-dessus qu'un extrait, assez pour comprendre qu'il est de peu d'intérêt, d'abord parce qu'il est trop détaillé, mais surtout parce qu'il ne rend pas justice au caractère ordinal et métrique des variables numériques, qu'il traite juste comme des nominales, alors que le résultat graphique est beaucoup plus approprié.

Pour ces raisons, le résultat tabulaire est rarement utilisé dans le cas numérique × numérique.

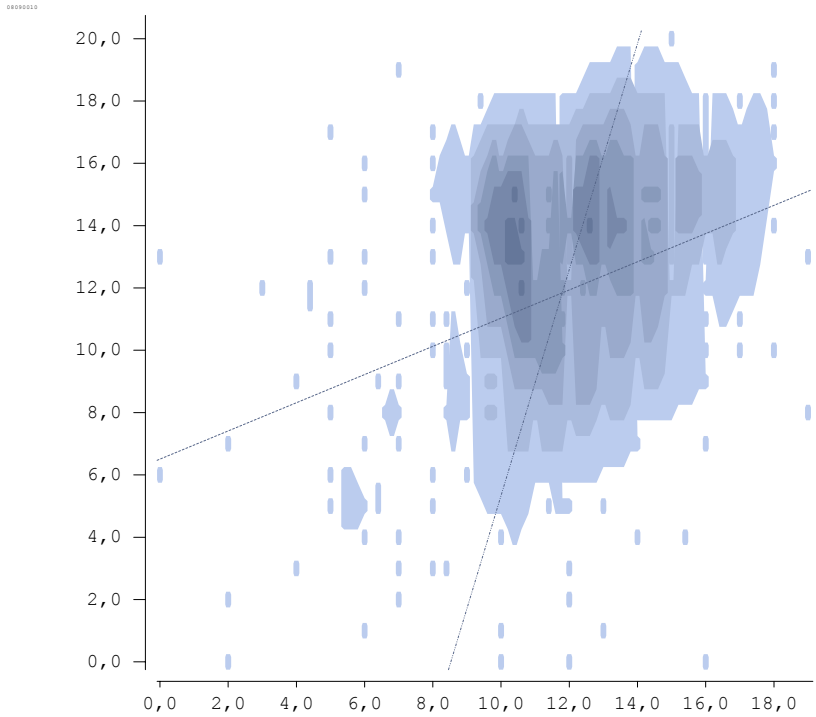
En revanche, la méthode exposée antérieurement reste valide : on examine la statistique globale :

```
r (Bravais-Pearson) = 0,353 , s. à .01
```
[129]

Dans ce cas, c'est le coefficient de corrélation de Bravais-Pearson souvent noté r_{BP}. Ce coefficient symétrique[130] se calcule selon une formule présentée dans l'annexe statistique. L'important est de savoir qu'il mesure la tendance des deux variables à « varier ensemble », c'est-à-dire que les grandes valeurs de l'une se trouvent chez les mêmes sujets que les grandes valeurs de l'autre, et de même pour les valeurs faibles.

Le résultat graphique ci-dessus peut aider à comprendre ce qu'est la corrélation.

Ce *graphe* est en *nuage de densité* : il représente des groupes de sujets par des taches d'autant plus sombres que les sujets concernés sont nombreux. Les valeurs des deux variables servent de coordonnées cartésiennes dans le plan pour situer les taches.



Le trait en pointillé large, proche de l'horizontale, est la droite des moindres carrés de la régression d'y en x, c'est-à-dire la droite qui représente le mieux l'idée que l'ordonnée y (coordonnée verticale) d'un sujet pourrait se calculer d'après son abscisse x (coordonnée horizontale) à l'aide d'une formule du genre y = ax+b, plus pas mal de *bruit* expliquant pourquoi tous les points ne sont pas sur la droite.

L'autre trait pointillé représente symétriquement la droite de régression d'x en y, selon une formule du style x = a'y + b'. Le coefficient de corrélation de Bravais Pearson

[129] s. à .0000 dans la version plus récente du logiciel.

[130] La corrélation de (x,y) est la même que la corrélation de (y,x).

entretient des relations étroites avec les coefficients a et a', pentes des droites de régression ; en effet, $r_{BP}^2 = aa'$.

Concrètement, rBP est un coefficient qui prend ses valeurs entre -1 et +1.

- à +1, c'est une corrélation positive parfaite : le ciseau des deux droites se resserre au point qu'elles sont confondues ensemble et avec la diagonale principale (bas-gauche vers haut-droite)
- à 0, c'est une corrélation nulle : les deux variables sont parfaitement indépendantes, et les deux droites sont perpendiculaires[131], la première horizontale, la seconde verticale
- à -1, c'est une corrélation négative[132] parfaite : le ciseau des deux droites se resserre au point qu'elles sont toutes deux confondues avec la seconde diagonale (haut-gauche vers bas-droit).

La plupart du temps, le coefficient de corrélation prend une valeur intermédiaire. Dans l'exemple, il est de 0,353, ou .353. Ce n'est pas une corrélation très forte, mais elle est très significative (à .01, version 2005, .0000, version 2011) parce qu'elle s'exerce sur des effectifs nombreux.

On peut noter que tandis que le χ^2 fournissait deux informations, sa force (la valeur du χ^2) et le seuil de probabilité associé, le r_{BP} en fournit trois : sa force (la valeur absolue du coefficient), son sens (le signe du coefficient) et le seuil de probabilité associé. Ce dernier, comme pour le χ^2, est fourni par une table spécifique, à n-2 degrés de liberté, où n est l'effectif.

Comment interpréter ce résultat ? Compte tenu de la valeur du r_{BP}, on peut rejeter l' Hypothèse nulle, et conclure au seuil P= .01 (ou .0000) à une corrélation positive, peu élevée[133] mais très significative[134], entre les variables A_NOTE et B_NOTE. Et après ? Ca ne signifie pas que A_NOTE soit la cause de B_NOTE. Simplement, elles ont une modeste tendance à varier ensemble, ce qui peut s'expliquer par le fait que de bons étudiants ont des bonnes notes partout, et que les mauvais n'en ont nulle part ; peut-être aussi ces deux notes correspondent-elles à des modes d'évaluation similaires, telles qu'une dissertation sur table, et que c'est peut-être l'habileté dans cet exercice qui explique la corrélation.

Une précaution qu'on n'a pas prise

Le coefficient de corrélation de Bravais-Pearson s'applique en principe au croisement de deux distribuions approximativement normales. Etait-ce bien le cas de A_NOTE et B_NOTE ?

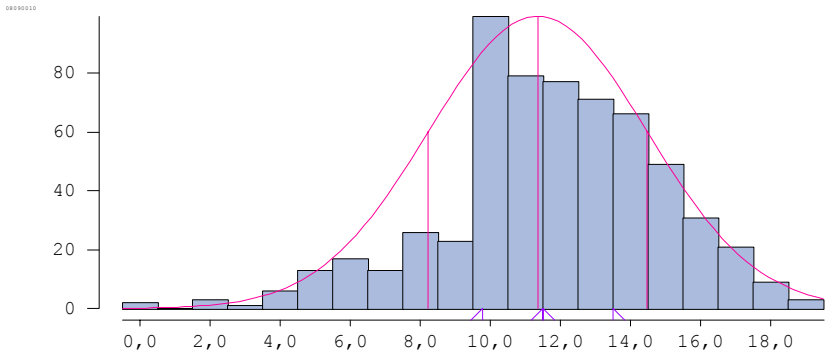
```
(A_NOTE)
 Valeur modale : 10,0 (n=84)
 Médiane entre 11,8 et 12,0
 Moyenne 11.80, écart-type 3.12
 H(normalité) rejetée à .10 ; H(symétrie) rejetée à .10
```

[131] Ou orthogonales, ce qui a souvent en mathématiques et en statistiques le même sens que indépendantes.

[132] Une corrélation négative concerne des phénomènes qui varient en sens inverse : exemple vos dépenses et ce que vous épargnez.

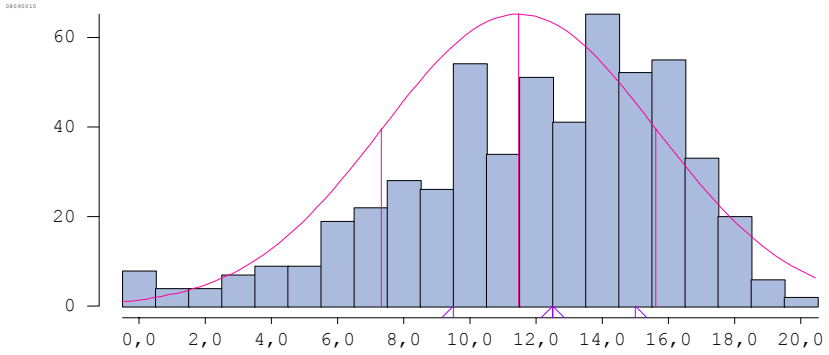
[133] Pour Piéron, fondateur de la docimologie, des valeurs élevées pour un coefficient de corrélation commencent vers .600 en valeur absolue.

[134] Encore une fois, à cause des effectifs importants (800 environ). Le même coefficient, avec des effectifs moindres, ne serait sans doute pas significatif.



Cette distribution n'est pas tout à normale, comme en témoignent les tests de normalité et de symétrie, dont les hypothèses sont rejetées à .10, ce qui n'est pas très grave mais tout de même. Qu'est-ce qu'elle a d'anormal ? Essentiellement le mode en 10 et la dépression en 9, ce qui typique d'une distribution censurée : il s'agit de notes pour des unités capitalisables à l'Université, et dans de nombreux cas où la note spontanée serait 9, un petit coup de pouce lié à l'incertitude ou à la mauvaise conscience docimologique la fait passer à 10. Si on imagine d'écrêter 30% de l'effectif de la note 10 et de le rabattre sur la note 9, la distribution devient quasi normale.

```
(B_NOTE)
 Valeur modale : 14,0 (n=62)
 Médiane entre 12,5 et 13,0
 Moyenne 11.93, écart-type 4.14
 H(normalité) rejetée à .0000 ; H(symétrie) rejetée à .0001
```



Le cas de B_NOTE est plus grave : les hypothèses de normalité et de symétrie sont rejetées à des seuils très fins, la distribution est en effet massivement dissymétrique à droite.

Alors ? On ne pouvait pas croiser ces deux numériques, parce qu'au moins l'une d'entre elles est gravement anormale ? Si, mais plutôt que le coefficient de corrélation paramétrique[135] r_{BP}, il faut utiliser son équivalent non-paramétrique, le coefficient de corrélation sur les rangs ρ (prononcer rhô) de Spearman, qui se calcule de la même manière, mais sur les rangs de classement des sujets plutôt que directement sur les valeurs.

```
(A_NOTE) x (B_NOTE)
 r (Bravais-Pearson) = 0.353 s. à .0000
 rhô (Spearman) = 0.317 s. à .0000
```

La corrélation par rangs est en effet un peu inférieure à la corrélation par valeurs, que la non-normalité des distributions incitait à ne pas utiliser.

Concrètement, dans cet exemple, la différence n'est pas énorme, mais dans des cas extrêmes il convient quand même de prendre cette précaution de vérifier la normalité.

[135] Qui utilise moyenne et écart-type comme des résumés pertinents d'une distribution.

Une variable nominale, une variable numérique

C'est le cas mixte. Il vise à répondre à la question suivante : les catégories découpées dans la population par la variable nominale ont-elles des caractéristiques différentes quant à la variable numérique ? Par exemple, des groupes ethniques différents ont-ils des tailles différentes ? Ou encore, des ingénieurs issus de différentes écoles ont-ils des salaires de début différents ?

Des gens qui ont des tailles ou des salaires différents, il y en a tout le temps. Le problème est d'arriver à distinguer si les différences entre individus s'expliquent plus par le groupe auquel ils appartiennent que par des variations individuelles. Par exemple, tel individu du groupe Dinka est-il grand parce que les Dinka ont tendance à être grands, ou parce qu'il est grand pour un Dinka ?

Prenons un exemple plus proche dans notre corpus-exemple. La variable ACCES décrit en détail comment les individus ont pu entrer en licence (autre licence, bac+2 technique, bac+3 social ou santé, DEUG, divers diplômes, expérience professionnelle, accès inconnu). L'âge n'a pas, hélas, besoin d'être présenté.

Le tableau brut de ce croisement n'offrirait pas grand intérêt car il serait trop détaillé. On utilise plutôt dans ce cas, comme résultat tabulaire, le tableau des effectifs, moyennes et écart-types selon les classes de la variable nominale. Celle-ci est donc implicitement traitée en variable indépendante.

D'ailleurs, même si on essaye de placer d'abord l'âge en lignes, puis l'accès en colonne, le logiciel utilisé ici rétablit l'ordre nominale-numérique, parce que c'est la seule présentation intéressante, raison pour laquelle le cas numérique × nominale est automatiquement converti en nominale × numérique.

Analyse de la variance de (Age) selon les positions de (ACCES)

Classe	Effectif	Moyenne	Ecart-type
aut.lic	35	26.46	6.03
B+2 tk	99	27.91	4.77
B+3 ss	45	33.51	7.00
DEUG	405	26.31	5.63
divers	8	33.00	5.89
expe.pro	144	34.95	7.36
inconnu	231	30.75	7.19
ENSEMBLE	967	29.22	7.09

La démarche employée ici est souvent appelée Analyse de Variance, ou ANOVA[136]. La statistique globale associée est le F de Snedecor-Fisher. Cette mesure, dont le calcul en détail est expliqué dans l'annexe statistique et qu'une option de Nestor permet même d'afficher, vise à exprimer le rapport entre la variance *interclasse* et la variance *intraclasse*.

La variance interclasse exprime les différences qu'il y a d'une classe à l'autre, chaque classe étant considérée comme représentée par sa moyenne ; dans un de nos exemples, la variance interclasse exprime les différences que l'on trouve entre le Dinka moyen, le Mandingue moyen et le Yoruba moyen[137]. La variance intraclasse exprime le fait qu'à l'intérieur de chaque classe la moyenne n'est qu'un résumé, mais qu'il peut subsister d'importantes différences interindividuelles ; dans le même exemple, la variance intraclasse exprime le fait que parmi les Dinka, il en est qui sont grands pour un Dinka ou plutôt petit pour un Dinka, et pareil parmi les autres peuples.

---

[136] Pour les anglophiles acharnés : ANalysis Of VAriance.

[137] Honnêtement, je n'y connais rien en ethnies africaines, et je ne sais même pas si celles que je nomme vivent dans la même région. Je crois seulement que ce sont bien des noms de peuples, et je présente mes excuses à tous en cas d'erreur : c'est juste pour construire un exemple.

L'idée de base est que si la variance interclasse est beaucoup plus grande que la variance intraclasse, il est clair que j'ai le droit de considérer que les différences entre individus s'expliquent essentiellement par leur appartenance de classe, et je prouve donc au passage qu'il existe des différences en quelque sortes structurelles entre les classes ou groupes, du genre : les Dinka sont plus petits que les Yoruba, etc …

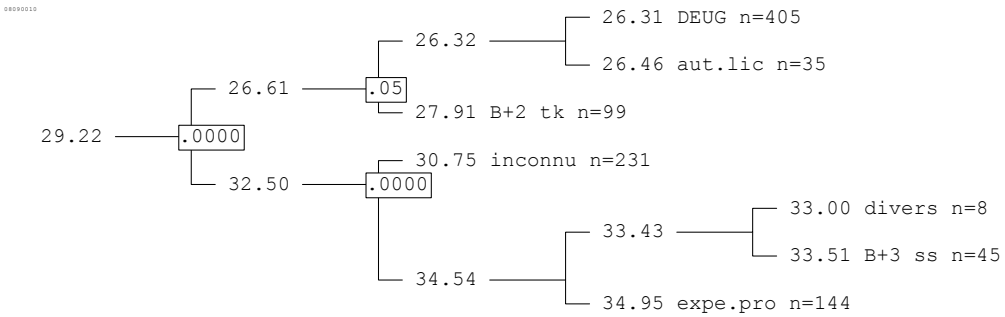
Si au contraire la variance interclasse est petite ou du même ordre de grandeur que la variance intraclasse, les différences entre groupes ne se dégagent pas du *bruit* causé par les différences entre individus, et il n'est pas légitime d'attribuer ces différences à l'appartenance de classe ; on s'en abstient donc : si le sujet 13 est plus grand que le sujet 44, ce n'est pas parce 13 est Dinka et 44 Yoruba, mais parce que 13 est une grande perche, et 44 plutôt râblé. Je ne peux rien conclure sur la généralité des groupes.

Dans notre exemple, la statistique globale est la suivante :

```
F(6,960) = 41.79, s. à .0000
```

Un F de Snédecor-Discher (à 6 et 960 degrés de liberté) de 41.79 est extrêmement significatif. Nous pouvons donc poursuivre l'analyse et commenter les résultats. Le tableau des moyennes nous donne bien des indications qui peuvent nous permettre de classer les modalités d'accès par âge moyen, mais ça ne nous dit pas si toutes les différences d'âge sont de même importance.

L'outil qui va nous informer sur ce point est la statistique locale du cas mixte, *l'arborescence des contrastes* :



C'est un drôle d'arbre, planté à gauche et couché sur le côté. La racine, 29.22, représente l'ensemble des 967 sujets, et 29.22 est, comme on le voit dans le tableau, la moyenne d'âge de l'ensemble.

A partir de cet ensemble qui comporte 7 groupes, le logiciel a commencé par classer ces groupes par moyennes croissantes ; puis il a essayé toutes les possibilités de dichotomie, c'est-à-dire toutes les façons de répartir les 7 groupes en deux paquets distincts. Pour chacune de ces dichotomies, il calcule le |t| de Student sur échantillons indépendants[138], et conserve la dichotomie qui amène le plus fort |t|. Ce |t| joue un rôle analogue au F de Snédécor-Fisher, mais pour comparer des groupes deux à deux alors que F, plus général, compare un nombre quelconque de groupes. Donc la dichotomie qui est conservée est celle qui établit le plus fort *contraste* entre les deux paquets.

Le paquet du haut, qui a pour âge moyen 26.61, comporte les groupes *DEUG, aut.lic et B+2 tk*, de manière indifférenciée dans un premier temps, tandis que le paquet du bas, avec 32.50 de moyenne, rassemble les groupes *inconnu, divers, B+3 ss et expe.pro*, également de manière indifférenciée dans un premier temps. Le rectangle superposé à la bifurcation qui sépare ces deux paquets contient l'expression .0000, qui témoigne que le |t| correspondant à cette dichotomie, à ce contraste, est significatif au seuil P = .0000.

On pourrait s'arrêter là, et ce serait déjà bien intéressant de savoir comment organiser les groupes en deux paquets distincts de manière à obtenir le contraste maximum, mais tant qu'à faire, le logiciel continue : il soumet chacun des deux paquets au même traitement

[138] Voir l'Annexe statistique.

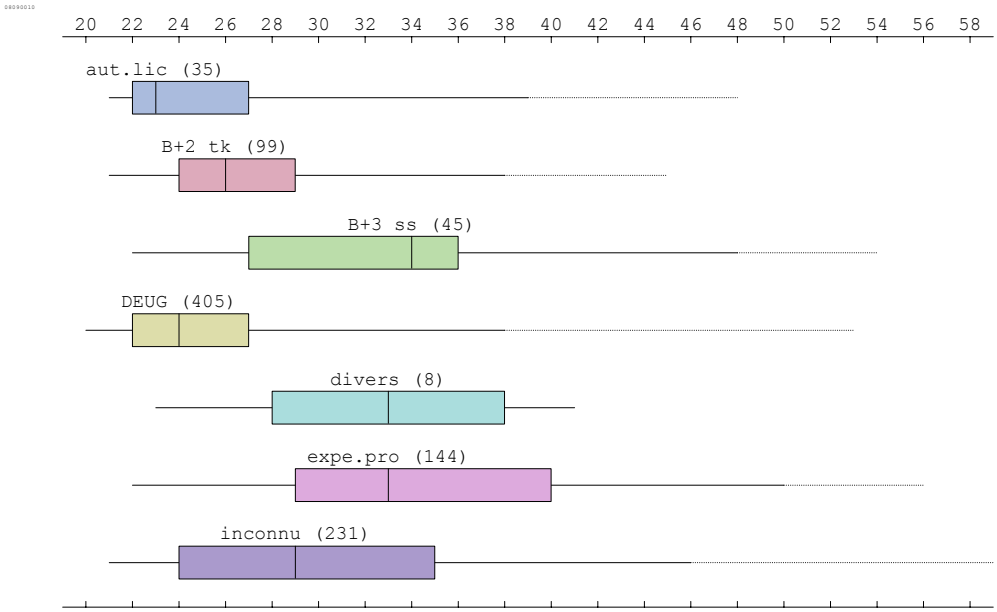
qu'avait subi l'ensemble, et cherche à créer des sous-paquets aussi contastés que possible. Et ensuite des sous-sous paquets, etc., jusqu'à arriver aux groupes d'origine qui sont comme les feuilles de l'arbre, et qui portent l'étiquette du groupe ainsi que son effectif.

Tant que le |t| qui arbitre le contraste est significatif, il est affiché à la bifurcation : si rien n'est affiché, c'est que le contraste n'est plus significatif : il n'y a donc plus lieu de le commenter.

On commente l'arborescence de gauche à droite et dans l'ordre croissant du seuil de probabilité (.0000 d'abord, .05 ensuite …) : les deux premiers paquets *[DEUG, aut.lic, B+2 tk]* et *[inconnu, divers, B+3 ss, expe.pro]* présentent un contraste très significatif ; puis le paquet du bas est découpé par un contraste très significatif entre *[inconnu]* et le reste. Dans le paquet du haut, un contraste secondaire à .05 oppose *[DEUG, aut.lic]* à *[B+2 tk]*.

Et puis c'est tout : les contrastes ultérieurs ne sont plus significatifs. En résumé, on obtient un découpage en trois sous-ensemble si on s'en tient au seuil de .01, en quatre si on accepte le seuil de .05 (et donc 5% de risque d'erreur). A l'intérieur de ces sous-ensembles, les moyennes d'âge ne peuvent pas être considérées comme statistiquement différentes.

On peut ensuite, si cela présente un intérêt, utiliser le résultat graphique pour illustrer le commentaire. Le graphe du cas mixte est le schéma en *boîtes à moustaches* :



Les échelles identiques en haut et en bas rappellent sommairement les valeurs de la variable numérique (ici *Age*). Chaque machin de couleur correspond à l'une des catégories, ou groupes, créés par la variable nominale (ici *ACCES*). Le machin, qui est une boîte à moustaches[139], se décrit de la manière suivante :

- de l'extrémité gauche à l'extrémité droite des moustaches, c'est l'étendue totale des valeurs de la variable numérique pour ce groupe de sujets.
- si une partie de la moustache est en pointillés, c'est qu'elle s'étend au-delà de deux écarts-types de la moyenne, et qu'il s'agit donc de mesures rares et isolées : autrement dit, on n'est plus dans la partie utile de la distribution, et il n'y a pas grand inconvénient à ne pas tenir compte de ces valeurs

[139] Ce nom, qui n'est pas un trait d'humour potache de la part de l'auteur, est la traduction littérale du *box and whiskers* de la très sérieuse littérature statistique anglo-saxonne (et on sait que ces gens ne rigolent pas tous les jours). Il ne faut pas l'entendre comme une boîte où l'on rangerait ses moustaches, postiches à l'évidence, mais comme une boîte prolongée latéralement par des moustaches.

• la boîte en couleur est coupée par un trait vertical qui représente la médiane des valeurs pour cette catégorie, c'est-à-dire la coupure telle que le nombre de sujets de part et d'autre est approximativement à 50/50%

• entre le bord gauche de la boîte et la médiane, il y a approximativement 25% des sujets[140], et de même entre la médiane et le bord droit de la boîte

• il y a donc à peu près la moitié des sujets dans la boîte, et un quart dehors de chaque côté.

Dans certains cas, ce dispositif visuel est très précieux pour comparer rapidement les différents groupes. Ne pas perdre de vue toutefois que les commentaires doivent s'appuyer sur les statistiques locales, après le feu vert de la statistique globale[141].

Synthèse : résumons-nous

Les tests et graphiques proposés peuvent posséder de nombreuses variantes, et même permettre de travailler sur des relations assez différentes si on réintroduit dans le débat les logiques et les ordinales.

Toutefois, les trois piliers de la démarche statistique inductive sont là, et connaissent des prolongements dans les techniques d'analyse de données les plus complexes : le croisement de nominales débouche sur l'Analyse Factorielle de Correspondances Multiples, le croisement de numériques est la base de l'Analyse en Composantes Principales[142], le croisement mixte se prolonge avec l'Analyse de Variance Multiple[143].

Ce qu'il est impératif de retenir de ce trop bref exposé tient en peu de mots :

• On examine toujours en premier la statistique globale

• Si la statistique globale n'est pas significative, on en prend acte et l'on s'abstient de commenter les particularités du tableau

• Si la statistique globale est significative, on s'appuie sur la statistique locale[144], quand elle existe, et/ou sur le graphe pour commenter la relation attestée.

Le tableau suivant résume les trois cas de figure principaux et leurs caractéristiques :

Cas	Tableau	Stat. Globale	Stat. Locale	Graphe
Nominale × nominale	De contingence	χ^2	% lignes, signe des associations locales	AFC simple, Barres et pourcentages
Numérique × numérique	Peu utile		*aucune*	Nuage de densité Droites de régression
Nominale × numérique	Des moyennes et Ecarts-types	F	Arborescence Des contrastes	Boîtes à moustaches

[140] Ces quarts d'effectifs approximativement égaux s'appellent des quartiles.

[141] Il peut se produire qu'à un F non significatif soit associé une arborescence comportant au moins un contraste plus ou moins significatif. Cela ne signifie pas qu'un test dément l'autre, mais que si on regroupait la variable pour en faire une opposition simple selon le contraste proposé, on obtiendrait quelque chose de significatif. Mais, alors, il ne s'agirait plus de la même variable, puisque sa structure aurait changé. La significativité du |t| est en quelque sorte relative à une ou des variables virtuelles.

[142] Mais aussi de la très grande famille des techniques basées sur l'analyse des corrélations, comme la régression multiple les modèles d'équations structurales ou analyse cheminatoire.

[143] MANOVA chez les anglo-saxons : une variable dépendante numérique, plusieurs variables indépendantes nominales ; cette technique permet de déterminer l'influence respective de plusieurs facteurs de variation.

[144] Exception faite de l'arborescence des contrastes qui n'est pas véritablement une statistique locale, mais une série de statistiques globales sur des regroupements de valeurs, et qui peut donc être commentée même si le F n'est pas significatif.

Pour les curieux

Le logiciel utilisé pour les exemples, qui est à la fois un outil pour le chercheur et un instrument de la didactique des statistiques, propose pour cette dernière raison des choix d'outils par défaut dans toute situation, pour éviter au débutant l'embarras de fouiller dans une boîte d'outils dont il ignore l'usage. Les cas particuliers ne sont pas pour autant ignorés. On en donne ici une liste rapide. Les personnes intéressées iront chercher plus de détails dans la documentation :

Dans les croisements de numériques, il est possible d'obtenir, optionnellement :

- **reg** : Les coefficients de régression linéaire d'une variable en l'autre
- **rhô** ou ρ de Spearman : le coefficient de corrélation sur les rangs, à utiliser quand les distributions étudiées s'écartent trop du modèle de la Loi Normale, ou que leur qualité métrique est douteuse : on nie alors leur caractère de mesure pour ne conserver que leur caractère ordinal
- **|t|**$_a$: le |t| de Student sur échantillons appareillés, utile quand les deux variables expriment un couple de mesures avant et après, et qu'on cherche à mettre en évidence un effet de progrès.

Le croisement de nominales autorise, outre le χ^2, **cn**, coefficient normé de contingence, mesure dérivée du χ^2, mais qui offre l'avantage d'être indépendante de la forme du tableau et de l'effectif ; valant entre 0 et 1, ce coefficient permet donc de comparer la force de la relation de contingence entre tableaux différents.

Les croisements d'ordinales autorisent :

- le χ^2 et le **cn** comme les nominales
- mais aussi le **rhô** ou ρ de Spearman comme les numériques, à cause de l'ordre
- et possèdent un test spécifique : **gGK**, le γ de co-ordonnancement de Goodman-Kruskal

Les croisements de logiques autorisent :

- le χ^2 et le **cn** comme les nominales
- le r_{BP} et le **rhô** ou ρ de Spearman comme les numériques
- le **gGK**, comme les ordinales
- un test spécifique, l'implication **=>**
- les statistiques de prédictibilité, sensibilité, spécificité

Devant une numérique, les variables logiques et ordinales se comportent comme des nominales, et offrent les résultats du cas mixte : tableau de moyennes, F, arborescence et boîte à moustaches. On peut aussi obtenir le détail du calcul du F, qui est apprécié dans certaines publications, avec notamment la statistique η^2, interprétée comme un taux de variation expliquée par la ou les variables indépendantes.

Croisées entre elles, logiques, ordinales et nominales (par ordre de vertu métrique décroissante), se comportent comme des variables de la plus basse vertu : devant une nominale, une ordinale oublie qu'elle est ordinale et agit comme une nominale.

Pour conclure

Rappel à la modestie : ce qui précède n'est nullement l'embryon d'un ouvrage de statistiques pour les Sciences Humaines et pour l'Orthophonie : il en existe d'excellents, quoiqu'ils ne le soient pas tous. Cette reprise et réécriture partielle d'un cours d'initiation statistique trouve sa place ici comme préalable éventuel ou simple rafraîchissement de mémoire avant l'étude du guide proprement dit. Il y a beaucoup plus de choses intéressantes à faire avec les statistiques que ce qui est dit dans ces quelques pages, mais l'expérience montre à l'évidence qu'on ne va pas très loin dans l'étude de la statistique en soi, et que la meilleure façon d'apprendre est de pratiquer sur ses propres données.