Gilles HUNAULT

2000

Analyse des petits Systèmes Statistiques

Université d'Angers

"Il est en effet clair que, sauf pour qui pense que le commentaire de bistrot ou la conversation entre amis constitue *l'essence même de la vie politique*, la discussion n'est politique qu'autant qu'elle se cristallise dans une décision."

"Aucune séquence politique véritable n'est représentable dans l'univers du nombre et de la statistique."

Alain Badiou, 1998.

Abrégé de métapolitique
Editions du Seuil, coll. l'Ordre philosophique.

Table des matières

Chapitre 1.

Introduction

1.1 Notion de système d'analyse statistique

Les probabilités et les statistiques sont deux domaines complémentaires des mathématiques. Les probabilités fournissent le cadre théorique, indépendamment de toute réalité. Les statistiques partent de la réalité (ou plutot de données tirées de "la" réalité) pour donner une vue globale, tentent de synthétiser ou de modéliser en se servant des modèles théoriques issus des probabilités. Les statistiques élémentaires ont pour but de décrire les données, de les résumer, si possible sans trop les déformer. Le recueil, la collecte des données, notamment dans le cas d'enquête est une partie délicate et souvent négligée. Pourtant elle conditionne le reste des traitements. Il faut donc avoir des connaissances en méthodologie d'enquête pour mener à bien l'élaboration du questionnaire, la conduite d'entretiens...

En particulier les notions de protocole de passation (par enquêteur, au téléphone, par envoi de courrier...), de protocole de déroulement (par exemple en double insu, avec placebo pour les enquêtes médicales, pharmaceutiques...), de technique d'échantillonnage (par quotas, aléatoire...), de découpage en classes doivent être définis avec soin. Ce n'est pourtant malheureusement pas ni le lieu ni le temps de faire ici un tel cours et nous nous bornons donc souligner à ces problèmes, pensant qu'être sensibilisé(e) pousse parfois à la connaissance... On trouvera par contre dans la bibliographie les références nécessaires pour approfondir ce sujet.

Si en probabilités on parle d'épreuve, d'univers ou d'espace d'épreuves puis de valeur réalisée et de variable aléatoire, en statistiques on traite des *populations*, dont une partie nommée *échantillon* est analysée. Les constituants de la population, analogues aux évènements des variables aléatoires sont les *observations* nommées aussi *caractères* ou *variables* effectuées sur des *individus*.

Contrairement aux probabilités, les statistiques ne traitent pas forcément des variables aléatoires mais plutôt des *données*. Celles-ci sont souvent présentées en lignes ou en colonnes, dans des tableaux de valeurs. Traditionnellement, on met en lignes les *individus* et en colonnes les *variables statistiques*. On appelle aussi les colonnes *séries statistiques* parce que ce sont des suites ordonnées de valeurs. Les données correspondent le plus souvent à un échantillon, ce qui signifie qu'on n'a pas à notre disposition l'ensemble des valeurs pour la population complète mais seulement une partie de ces valeurs. La représentativité de l'échantillon est là aussi un problème de fond. Nous renvoyons à des ouvrages spécialisés pour ce problème. Mais les journaux (télévisés, de la presse écrite...) fournissent aujourd'hui les techniques (sondage par quota, par exemple), et les quantités utilisées (enquêtes dites "nationales" de 1 000 ou 2 000 personnes...) pour que le lecteur en ait quand même une petite idée.

Lorsqu'on traite des séries statistiques, on est le plus souvent confronté à des fichiers de données. Une des premiers tâches concrètes est de vérifier l'intégrité des données (aucun age n'est négatif ou nul...), leur cohérence (toutes les séries ont le même nombre de valeurs...), leur exhaustivité (les unités, les codes utilisés sont explicitement décrits). Contrairement aux probabilités où par exemple la loi binomiale peut être traitée sans aucune interprétation physique, les séries statistiques doivent être soigneusement décrites. C'est souvent une difficulté pour un spécialiste des calculs que d'être au fait des modalités de traitement, de comprendre ce que signifient les termes employés.

Ainsi, une analyse nommée $ANTAL$ vient présenter des données qui mettent en jeu des médicaments antalgiques. Certaines valeurs sont issues de traitements placebo en double-insu. Pour le statisticien, ces qualificatifs doivent avoir un sens, sinon la rédaction risque d'être incorrecte. De même, le dossier ELF (Enquête Linguistique sur la Féminisation des noms de métier) traite de l'*actio-régionnalité*. Comment interpréter une moyenne d'age de 25 ans par rapport à cette actio-régionnalité si on ne connait pas ce mot ? C'est là tout l'art et le travail du statisticien que d'apprendre, de survoler un domaine pour ensuite aider les spécialistes de la discipline à éclairer les données par des conclusions pertinentes...

Une étude statistique élémentaire vient distinguer les variables et traite différemment les variables qualitatives, les variables quantitatives et les variables textuelles. Après une partie dite d'analyse univariée (ou analyse à une dimension) et une autre d'analyse bivariée, (ou analyse à deux dimensions), on doit archiver les fichiers de résultats, commenter et mettre en forme les analyses et commentaires pour ensuite les imprimer, les exposer...

Un *petit système d'analyse statistique* est un système logiciel capable de réaliser toutes ces tâches, de la saisie des données à leur décodage, de leur analyse statistique à leur exposition, de la gestion des données à la gestion des fichiers. Ce cours vise à montrer comment utiliser des logiciels classiques orientés plutôt vers le calcul, la gestion pour en faire des petits systèmes d'analyse statistique, ce qu'ont doit attendre des logiciels statistiques et comment les intégrer dans une démarche exhaustive et critique.

La table des matières des pages précédentes a cité ces logiciels. Ajoutons seulement que fidèle à nos habitudes nous présenterons des solutions logicielles multi-plateformes (*Dos*, *Unix/Linux*, *MacOs*, *Mvs*, *Vm/Sp*...), des conseils pour s'adapter à "l'air du temps" qui vient dans le changement et la mise à jour commerciale compliquer la science statistique par des considérations mercantiles et technologiques.

Pour terminer cette première partie de l'introduction, reprenons notre *Credo du Statisticien* à réciter chaque soir après la prière :

- La machine calcule, l'humain interprète et rédige.
- Après et avec les calculs, les graphiques.
- Si les données sont fausses, les résultats sont faux.
- Si les formules sont fausses, les résultats sont faux.
- Si on emploie mal l'ordinateur, les résultats sont faux.
- Si les conditions d'expérimentations sont omises, l'analyse est sans doute incorrecte, pour le moins incomplète.
- Les résultats peuvent être justes, mais les interprétations fausses ou mal exposées.
- La statistique est une *science* qui requiert éthique, honnêteté, rigueur et volonté de communiquer ; l'ordinateur doit être un *outil* au service des hommes.

1.2 Statistiques et Analyse des Données

Après l'analyse à une et à deux dimensions, on pourrait penser passer à l'analyse à trois dimensions. En fait, ce n'est pas aussi simple. Passer de l'ordre deux à l'ordre trois n'est pas chose aisée. Tout le monde se souvient de la résolution de l'équation du second degré, mais de celle du troisième? On apprend jusqu'en terminale à tracer des courbes sur le modèle $y = f(x)$ soit deux variables en tout (x et y), mais tracer à trois variables suivant le modèle $z = g(x,y)$ est beaucoup plus ardu. Le même problème se pose pour la représentation : la dimension deux correspond au plan, aux courbes, la dimension trois à l'espace et aux surfaces. C'est pourquoi la statistique classique s'est longtemps arrêtée seulement à des généralisations vectorielles des notions précédentes avec des calculs "honnêtes" (pour un mathématicien), effroyables pour le commun des mortels comme la réduction des formes quadratiques définies sur des vecteurs gaussiens à p dimensions (pour $p >2$). En fait, on passe directement de 2 dimensions à n dimensions, pour toute valeur de $n >2$.

Avec l'analyse exploratoire des données, la statistique descriptive multidimensionnelle, les statistiques élémentaires ont trouvé d'autres champs d'application : la projection d'éléments dans des plans (factoriels) de plus grande inertie, la classification d'eléments en groupes (classes), le classement (affectation) d'éléments à des classes prédéfinies,.. Ces méthodes et techniques, qui ne requièrent aucune hypothèse sur les données sont désignées en France par le terme générique d'*Analyse des Données*.

Si on essaie de traiter simultanément un ensemble d'individus et de variables, un pré-requis est de disposer d'un tableau de données rectangulaire à n lignes et p colonnes pour que toutes les lignes appartiennent au même espace de représentation et de même pour les colonnes. Là où les statistiques classiques ne traitent que les colonnes, l'*Analyse des Données* vient travailler dans les deux dimensions (lignes **et** colonnes). De plus elle cherche des relations linéaires générales (à $v \leq min(n,p)$ composantes). Les méthodes sont différentes suivant qu'il s'agit d'un tableau de variables quantitatives (ACP), d'un tableau avec une seule variable quantitative ventilée suivant plusieurs critères (AFC), d'un tableau de contingence (AFC), de variables logiques, (AFC), de variables qualitatives (AFM) mélange de variables (AFM) mais il s'agit d'une même démarche nommée *Analyse Canonique*. On cherche des directions d'allongement de l'inertie ou information contenue dans le tableau des données, on projette et on représente sur les axes principaux d'inerties les éléments, lignes et colonnes.

Une suite logique de ces calculs est alors les méthodes de classification automatique hiérarchique, ascendante (CAH) ou descendante, prenant pour entrées soit des données brutes, soit des coordonnées sur des axes factoriels. Ces méthodes participent du même but que les statistiques descriptives élémentaires, à savoir décrire synthétiquement les lignes et les colonnes, faire entrevoir les liaisons entre ces éléments, les quantifier et en donner des représentations graphiques. Toutefois, pour des grands tableaux de données, on viendra classifier directement les données (et non pas leur coordonnées pondérées dans des espaces simplifiés de représentation) avec des méthodes comme les *Nuées Dynamiques*, les *Boules Optimisées*, l'*Agrégation autour de Centres Mobiles*...

Notre but ne sera pas d'enseigner ces méthodes car cela dépasserait le niveau d'inititation imparti à ce manuel, mais nous présenterons ces méthodes sur des exemples, afin de montrer comment on les utilise, ce qu'ils mettent en oeuvre, tant au point de vue statistique qu'informatique. De plus ces méthodes sont souvent, contrairement aux statistiques classiques, couplées à des sorties graphiques et il est très intéressant de savoir que ces méthodes existent, même si ce n'est qu'à travers des exemples qu'elles sont appréhendées...

1.3 Exemples d'Analyses des Données

Pour notre premier exemple, nous utiliserons un dossier nommé Logement qui fournit pour 36 lieux (des arrondissements de Paris, des grandes villes, des stations balnéaires, hivernales...) les prix minimum et maximum de logement au m^2 pour des années choisies entre 1970 et 1988. L'analyse effectuée est nommée AFC, ce qui signifie *Analyse Factorielle des Correspondances*, la correspondance étant ici l'application qui au couple (*lieu*,*an*) associe le prix "correspondant". L'AFC effectue une décomposition de l'inertie au moyens de valeurs propres et vecteurs propres calculés sur une matrice de distances issue des données et vient projeter les éléments de départ dans la base des vecteurs propres normés. Le graphique qui suit, nommé plan factoriel donne une représentation spatiale des lieux si on utilise les deux premiers vecteurs propres.

Ce vocabulaire technique ne doit pas inquiéter les non-mathématicien(ne)s : c'est l'ordinateur qui effectue tous les calculs, l'utilisateur n'a qu'à entrer le nom du dossier et consulter le fichier de sortie.

```
+-------------------baul-----+------------------------+
!  NICE ANTB                 !                        !
!                            !            pa14    pa15!
CANN                         !                 pa12   !
!                            !       PA16             !
!                            !  tolo      pa05    pa07!
!                            !                        !
!                megv        !   bord                 !
!                            !                    stlr!
+----------------------------+------------------------+
!                    lgrm    !                        !
!            nant            roya                 PA20!
!    aixp                    !                    CRCH!
!        deau        arca    ar/m            vldi     !
!                stra    lyon!                        !
!                            !   mars                 !
!                            mntp                 LROU!
!    stmx                prpg!                        !
!                fnrm        !                        !
!        alph                !      AVOR              !
+----------------------------+------------------------+
```

Ce plan factoriel met en évidence, soulignée par nous avec des lieux en majuscule, une opposition sur l'axe un (horizontal) entre `{CANNes,NICE,ANTiBes}` et `{PAris 20`^{e}`,CouRCHevel}` et au niveau de l'axe deux (vertical) une opposition entre `{AVORIAZ,LaROUsse}` et `PAris 16`^{e}.

On se méfiera de ces oppositions, de même de que les attirances (CANNes et NICE) par exemple ne peuvent pas être lues directement mais sont déduites des sorties numériques associées à l'AFC car ce ne sont pas les distances (visibles sur le graphique) qui définissent ces relations mais des indicateurs pondérés (et on ne voit pas les poids des éléments sur le graphique) basés sur les inerties des éléments (l'inertie étant une masse multipliée par une distance au carré, elle n'est pas non plus accessible directement sur le graphique).

De même, les oppositions sur l'axe un sont plus fortes que sur l'axe deux, leur poids respectif étant défini par la valeur propre considérée... On comprendra sur ces remarques que si la pratique des logiciels d'Analyse des Données ne demande en informatique que de savoir exécuter un programme, elle requiert, par contre, quant au "dépouillement" des listings de sortie, quelques précautions et un bon entrainement pour une rédaction fine et pertinente.

Notre deuxième exemple sera un exemple de sortie en CAH qui est l'acronyme de *Classification Ascendante Hiérarchique*. Contrairement à l'AFC qui oblige à regarder de nombreux plans factoriels, la CAH vient donner sur un seul dendrogramme (ou arbre de classification) les regroupements deux à deux des données ou *agrégations binaires*, ce qui permet ensuite de définir une *partition* en 2,3...k classes. Pour cet exemple, nous utiliserons le dossier VINS dont les données sont des importations de vins (en milliers d'hectolitre) pour un certain nombre de pays suivant diverses catégories de vins (`CHaMPagne,vin d'ALSAce`...). La classification, traduite par la représentation graphique suivante

```
RHON *----*-----*-----------------------------------*
     !    !     !                                   !
GIRO +    !     !                                   !
     !    !     !                                   !
VDQS +    !     !                                   !
          !     !                                   !
PROV *----+     !                                   !
     !          !                                   !
BOJO +          !                                   !
     !          !                                   !
AOCX +          !                                   !
     !          !                                   !
XXXF +          !                                   !
                !                                   !
XXFF *----------+                                   !
     !                                              !
ANJO !                                              !
     !                                              !
MUSC !                                              !
     !                                              !
AOCF +                                              !
                                                    !
MOSE ---*--*----------------------------------------+
        !  !
CHMP ---+  !
           !
ALSA ------+
```

permet de distinguer, pour nos données deux classes d'importation, l'une constituée des vins `{MOSElle,CHaMPagne,ALSAce}`, l'autre avec les vins restants.

Une classification de type CAH utilise une *matrice de distances initiales*, un *critère d'agrégation*, une *formule de recalcul de distance* entre les anciennes classes et la nouvelle classe formée. Par exemple, le critère de $Ward$, basé sur l'inertie, permet d'effectuer une classification sur des coordonnées factorielles pondérées qui fournit une suite logique à l'AFC. Pour d'autres types de classification, on utilise d'autres critères comme le *single linkage criterion* (nommé aussi "saut minimum" ou encore "diamètre minimal"), le *"complete linkage criterion"* (saut maximum...), d'autre formules de recalcul comme l'*UPGMA* ou *unweighted pair group method (with) averages...*

D'autres classifications ne sont pas ascendantes mais descendantes, certaines ne sont pas binaires, d'autres encore produisent une partition au lieu d'une hiérarchie, d'autres enfin fournissent des classes floues, emboitées, des pyramides... La liste de ces termes [1]montre la diversité à défaut de la richesse des ces méthodes.

Un système complet d'analyse statistique se doit de comporter toutes ces méthodes et techniques, avec leurs options, leurs raffinements. Un *bon* système d'analyse statistique doit comporter en plus un système d'aide à l'utilisation, des garde-fous contre une utilisation aveugle (effectuer une AFC sur des données postives et négatives ou lancer une classification de type CAH sur des données qualitatives codées a autant de sens que de faire la moyenne de numéros de téléphone), fournir des aides à l'interprétation, faciliter les tri des résultats comme les moyennes, variances, contributions (parts d'inerties), coordonnnées...

1. et encore nous ne rentrons pas ici dans discussion sur la "foultitude" des coefficients de [dis]similarité, distance, ressemblance, proximité, ni dans le choix d'une distance : Jaccard ? Jaccard-Sneath ? Czekanowski ? Kulczynski ? Russel-Rao ? Ochiai ? Kochen-Wong ? Sokal-Sneath ? Sokal-Michener ? Rogers-Tanimoto ? Hamman ? Roux I, Roux II ? (liste non exhaustive !).

Chapitre 2.

Données et Fichiers

2.1 Données : Codification

Commençons par distinguer trois types élémentaires de données : celles *à unités*, comme les *kilos*, les *cm* qu'on nomme données quantitatives de celles *à codes*, qu'ils soient numériques, caractères, ordonnés ou non, qu'on nomme données qualitatives et celles *à mots*, correspondant à des valeurs stockées en tant que phrases et textes, qu'on nomme textuelles. Nous verrons au chapitre suivant que cela aboutit aux trois types de variables statistiques nommées variables quantitatives, variables qualitatives et variables textuelles. Il y a d'autres types de données que nous présenterons plus loin au chapitre 3.

Si le stockage de données quantitatives se fait en conservant les valeurs telles quelles, dans le cas d'un découpage en classe, il ne faut pas stocker le code de classe mais bien la valeur. Ainsi, pour des classes d'ages, il est bon de stocker l'age exact puis de calculer les codes de classes d'ages par programme. En cas de modification des choix de classes,[1] puisqu'on dispose des données de départ, on peut recoder. Par contre, les imprudent(e)s qui n'ont que les codes de classes sont obligé(e)s de repartir à "la pêche aux données".

Pour des données qualitatives, il est souvent suffisant de conserver les codes de modalités dans le fichier des données, à condition de garder dans un fichier annexe (souvent nommé fichier descriptif, de type .dsc) la correspondance entre code et label des modalités.

1. ce qui n'arrive **jamais** au moment de la saisie des données, mais **toujours** quelques mois plus tard, quand les fichiers initiaux ont disparu...

Ainsi, pour une variable a deux modalités codées 1 et 2 la liste des 1 et des 2 est claire, à condition que le fichier descriptif des variables n'ait pas été égaré. Dans certains cas, il faut recourir aux variables indicatrices des modalités. Ce sont l'équivalent des fonctions caractéristiques d'ensembles. Pour la variable qualitative considérée plus haut, les codes 1 et 2 doivent alors être respectivement remplacés par `1␣0` et `0␣1`. Le fichier des données qualitatives (ou quantitatives découpées en classes) décodées doit être vérifié avec soin pour être un fichier DBC (*disjonctif binaire complet*) acceptable. Le décodage des modalités en indicatrices assure la binarité et la disjonction mais pas la complétude. Cela signifie que certaines modalités peuvent ne pas être représentées suivant l'échantillon considéré. Ainsi, on peut avoir prévu 12 modalités sur l'ensemble des fichiers à traiter mais l'échantillon x peut ne mettre en jeu que les modalités 1 à 8. Les colonnes de décodage 9 à 12 seront alors de somme nulle, d'où des débordements de capacité en cas de calcul relatif (par exemple division par le total de la colonne en cours). De même, pour les *codages flous pondérés* (ou *codages probabilistes* pour lesquels on ne met pas un `1` et des `0` mais des nombres positifs de total 1), il est bon de vérifier que la somme des codages fait 1, que chaque code est recensé, présent au moins une fois...

Plus généralement, il faut toujours vérifier systématiquement les données, ne serait-ce qu'à cause des délimiteurs de décimales (le point et la virgule suivant les pays), mais aussi parfois l'espace ou la virgule pour les milliers, les puissances multiples de 3. De même, le codage des données accentuées, des fins de ligne sous *Dos*, *Windows*, *Unix*[2] peuvent parfois donner des valeurs numériques ou textuelles surprenantes. En particulier, les nombres comme .2 (acceptables pour les normes américaines) sont parfois rejetées comme incorrectes et doivent être complétées en 0.2 avant d'être reconnues.

Un tableur rajoute des difficultés supplémentaires :

- le *format de cellule*, qui ne montre pas la valeur réelle de la donnée mais une représentation formatée, par exemple 0.17 au lieu de 0.16985...,
- la *formule* qui n'est pas une valeur mais le résultat d'un calcul, interdisant par exemple des trier les cellules, obligeant alors à faire du *collage spécial des valeurs* avant de trier...

2. Une fin de ligne sous *Dos*, *Windows* se compose de deux caractères, `0x0D` ou "cr" (*carriage return*, retour charriot) et `0x0A` ou "lf" (*line feed*, saut à la ligne) alors que sous *Unix*, il n'y a que `0x0A`.

Pour ces tableurs, un coup d'oeil au cadrage droit ou gauche[3] permet parfois de prévoir tout de suite que le moindre calcul statistique aboutit inévitablement à un message du genre *Invalid Data Type*. En particulier, on prendra soin de noter que les logiciels statistiques sont *avides* et *stupides* et qu'ils ne peuvent s'empêcher de calculer la moyenne de colonnes numériques de données dès qu'on a le dos tourné, ce qui peut avoir des conséquences graves non sur le calcul mais sur la crédibilité, la rigueur de la personne ayant laissé passer cette erreur. On aura donc intérêt à utiliser des champs de type caractère plutôt que numérique pour stocker les codes des variable qualitatives, obligeant à noter `'1'` ou `"1"` plutôt que `1` la valeur correspondante, par exemple, pour faire ressortir ce typage ou utiliser un logiciel qui distingue les valeurs numériques qualitatives et les valeurs numériques quantitatives.

2.2 Fichiers : formats

En informatique, la représentation et le stockage des données, des résultats sont primordiaux. Il faut distinguer les formats de stockage, de représentation et d'affichage. Au niveau du stockage, de nombreux formats sont propriétaires, c'est à dire privés aux logiciels. D'autres sont déposés, publics, comme le format `.dbf` de *Dbase*.[4]

Un format intéressant est le format `.sdf` qui signifie *Standard Data Form*. Il correspond au mode texte simple (on dit aussi PRN ou "printer mode", mode texte ou TXT ou encore ASCII pur, même si tous les fichiers utilisent le codage ASCII[5]). Pour ce format, tout est bien cadré, sans caractère de controle. C'est souvent un format d'échange, en entrée (import) comme en sortie (export). C'est aussi un format long, puisqu'il utilise la taille maximale de chaque variable. La même remarque s'applique aux chaines de caractères dans ce format et oblige à travailler avec la longueur maximale des chaines.

3. Pourquoi ?

4. Ainsi on peut prévoir la taille d'un fichier *Dbase* : c'est $n_1 + n_2 + n_3$ (à quelques octets près) où n_1 vaut 32 octets pour l'en-tête (date, nombre d'enregistrements...), n_2 vaut, en octets, 32 fois le nombre de colonnes, n_3 correspond à la taille des données mises bout à bout). Un fichier *Excel* est 2, 3, 4, 5 fois plus gros, pour ces mêmes données, suivant la version d'*Excel*, le style de feuille ou de classeur...

5. En fait, IBM utilise aussi un codage EBCDIC ; le XXI^e verra surement l'utilisation de codages sur plus d'octets (l'ASCII en utilise 8) comme l'UNICODE (32 octets) qui permet des documents *multilingues*, les jeux de caractères (et non pas les polices de caractères) multiples...

Imaginons par exemple qu'on ait les valeurs "IDEN 0.217" et "Constitution 186.3" à stocker en format SDF, chacune sur une ligne. Cela s'écrit alors soit

```
IDEN␣␣␣␣␣␣␣␣␣␣␣␣0.217
Constitution␣186.3
```

soit

```
IDEN␣␣␣␣␣␣␣␣␣␣000.217
Constitution␣186.300
```

où le symbole ␣ représente ici un espace.

Un format tout aussi général mais plus court est le format .dlm ou *Delimited*. Comme son nom l'indique, les données ne sont plus formatées, cadrées, mais seulement séparées par un même symbole. Au format américain, la virgule est un délimiteur usuel. Si on ne traite que des valeurs numériques, l'espace est un bon séparateur. Pour un fichier contenant des données numériques et textes, le symbole # est souvent utilisé. Ainsi, les valeurs "IDEN", 0.217 et "Constitution", 186.3 à stocker en format .dlm, chacune sur une ligne, s'écrivent

```
IDEN#0.217
Constitution#186.3
```

Un bon logiciel doit avoir ces différents formats disponibles, en plus d'un format personnalisé et adapté au logiciel. Le choix de la taille des données, du nombre de décimales n'est pas aussi simple qu'il y parait. Certains logiciels, comme *Dbase* utilisent le format des données pour déterminer le format des résultats. Ainsi la moyenne de nombres entiers redonne un nombre entier, ce qui peut-être choquant si l'on n'est pas prévenu(e). De même, un format d'affichage comme sous Excel peut induire en erreur. En général, un format trop juste (comme 3 cases pour des nombres de 0 à 250) ne permet pas d'afficher de sommes, des moyennes, des pourcentages. On veillera donc à surdéfinir les formats d'affichage ou les cases des masques de saisie. Comme il n'est pas possible de prévoir à l'avance toutes les plages de variation, il faut très souvent se soucier de l'ordre de grandeur des données pour prévoir l'affichage.

On se méfiera aussi qu'un format numériques entier sur 3 caractères donne la plage de variation [-99,999] et non pas [000,999] ou [-999,999]. Un logiciel comme *SAS* sait distinguer les formats de sortie nommés simplement FORMAT des formats d'entrées INFORMAT, ce qui ne présage d'aileurs rien comme format de stockage.

L'exemple le plus intéressant à ce point de vue est celui des dates : on a souvent (par ce que c'est le plus court à écrire) des entrées comme 02/05/1999 affichées comme "2 Mai 1999" ou "Feb 2nd, 1999" mais souvent stockées autrement. Ainsi pour *SAS* une date est le nombre de jours écoulés entre le premier janvier 1960 et la date considérée.

2.3 Gestion des Fichiers d'une analyse

La gestion des fichiers relatifs à une étude statistique n'est pas une mince affaire. En plus du ou des fichiers de données en mode texte (.txt, .dat...), leur conversion en fichier *Dbase* (.dbf) ou *Excel* (.xls) suivant les besoins, les fichiers de calculs, de résultats, de commentaires, sous *Word* (.dbf), *Excel* (.xls), *LaTeX* (.tex, .dvi, .ps), ou autres, obligent à une gestion rigoureuse, une désignation standard des fichiers.

Il convient par exemple pour un *dossier* nommé ANTAL de nommer antal.dat le fichier des données, antal.dsc son descriptif, antal.dbf le fichier *Dbase* correspodant aux données etc. Cela permet, avec une notation ambigue comme antal*.* (sous *Dos*, *Windows*) antal* (sous *Unix*) de gérer l'ensemble des fichiers relatifs au dossier.

Suivant le nombre et la taille des fichiers, on viendra utiliser un répertoire commun aux analyses comme `k:\stat_ad\` en local ou `http://..~gh/wdata/` sur le *Web*, à moins qu'il ne soit plus intéressant d'utiliser un répertoire spécial pour chaque dossier comme `k:\stat_ad\antal`, `k:\stat_ad\elf`... Pour le stockage, l'archivage après l'analyse, il est bon de *compresser* et d'*archiver* avec des logiciels et commandes comme zip, gzip, pkzip, winzip ou *consor* comme compress, tar, wintar, arj, rar, stuffit... L'archivage regroupe en un seul (gros) fichier un ensemble de fichiers alors que la compression diminue la taille d'un fichier. Les logiciels sous *Dos*, *Windows*, *MacOS* réalisent souvent les deux opérations en même temps. Sous *Unix*, la compression suit souvent l'archivage. Ainsi avec la commande tar on produit un fichier .tar que la commande compress vient compléter en .tar.Z ou .tgz. L'intérêt de la compression et de l'archivage est énorme : un seul fichier à déplacer, transférer, copier avec des taux de transferts moindres puisqu'ils sont souvent liés à la taille du fichier. Un exemple flagrant : une image au format .bmp de 600 K se compresse[6] souvent en un fichier de 40 K, un fichier texte de 600 K devient un fichier compressé de 200 à 300 K en général...

6. c'est pourquoi les formats .gif, .jpg sont plus intéressants : ils sont déjà compressés !

Chapitre 3.

Variables et Traitements

3.1 Types de variables

Nous restreignons volontairement pour l'instant nos données à trois types élémentaires, les *variables à unités*, les *variables à codes* et les *variables à mots* nommées aussi variables quantitatives, variables qualitatives et variables textuelles. Les autres variables seront évoquées en fin de chapitre. Nous avons déjà discuté au chapitre précédent le problème du codage, recodage, formatage des données. Rappelons seulement qu'à coté du fichier des données, on doit avoir un fichier descriptif qui rappelle le type des données, les unités, les codes, voire le protocole de saisie, l'origine des données...

Nous noterons **QT** toute variable dont la ou les unités sont clairement définies et pour laquelle la notion de somme, de moyenne a un sens. Le mot QT est bien sur un acronyme pour (variable) QuanTitative. De même, nous désignerons par **QL** (variable QuaLitative) toute variable pour laquelle les valeurs (numériques ou caractères) n'on en fait aucune importance, parce qu'elles correspondent à un codage arbitraire. Enfin, nous désignerons par **QX** (variable teXtuelle) toute variable dont le contenu correspond à des mots, phrases ou textes.

Si les unités ne sont pas fournies pour une variable quantitative, il faut refuser d'effectuer l'analyse. On peut toujours croire être en mesure de les trouver, mais c'est un leurre.

Ainsi, l'age est une variable quantitative. Mais pour une réaction chimique, l'age des bactéries se compte en pico- ou en micro-secondes, pour un nourrisson, on utilisera des heures ou des jours, pour une étude géoclimatique, on pourra compter en mois ou en années... En ce qui concerne les codes (ou qualités, ou modalités) d'une variable qualitative, c'est encore pire : si au cours d'une enquête on demande aux personnes interrogées leur sexe, le code 0 peut être aussi bien celui de l'homme que celui de la femme. Qui plus est, le codage 1/2 (de la Sécurité Sociale) est national et ne sera donc pas employé par d'autres nations. Enfin, signalons un problème délicat, celui des non-réponses. Ainsi, pour une enquête récente, nous avons eu trois code-sexe : 0, rencontré 17 fois ; 1, rencontré 235 fois et 2, renconré 12 fois. Où sont les femmes ?

On prendra donc soin à bien référencer les variables, leur nature, les unités et les codes employés. Comme son nom l'indique, une base de données ne contient que les données. Il faut donc (et ce n'est pas un problème mathématique mais un problème logistique) lui adjoindre un ou plusieurs fichiers de description, voire un lexique des termes employés afin de savoir ce que l'on traite. D'où l'importances des fichiers nommés descriptifs.

Chaque variable, quel que soit son type, se traite à deux niveaux en statistique : séparément, comme si chaque variable était seule, puis conjointement avec chaque autre variable de même type, comme si on disposait de tableaux à deux colonnes. Le vocabulaire traditionnel nomme aussi analyse univariée ou analyse à une dimension ce que nous avons présenté comme l'analyse séparée et analyse bivariée ou analyse à deux dimensions notre analyse conjointe. On dispose bien sûr de termes plus spécialisés (comme *tri à plat*, *tric croisé*, *paramètres de dispersion...*, *lexique occurentiel*, *concordances...*) suivant le type de variable considéré.

3.2 Traitement des Variables quantitatives

Pour une variable quantitative, au niveau de l'analyse séparée, on effectue un calcul calqué sur le celui des variables aléatoires en Probabilités : moyenne, variance, écart-type etc. On attachera de l'importance à choisir le nombre de décimales d'affichage de telles valeurs, de façon à bien faire ressortir le phénomène de répartition des valeurs, que ce soit une équi-répartition ou une répartition inégale. On adjoindra aux calculs et indicateurs numériques des graphiques, courbes, représentations en étoile, diagrammes en "boites à moustaches"... pour montrer ce que les valeurs chiffrées indiquent...

Rappelons les paramètres de localisation (ou tendance centrale) et de dispersion (absolue, relative) d'une **QT** notée V avec n valeurs notées V_i pour i de 1 à n :

Paramètre	Notation	Formule
nombre d'observations	n	
moyenne	m	$(1/n)\sum V_i$
écart-type	σ	$(1/n)\sum V_i^2 - m^2$
coefficient de variation	cdv	$100.\sigma/m$

rappelons aussi que ces paramètres correspondents à des phénomènes :

Phénomène	Paramètre
Taille	nombre d'observations
Localisation	moyenne
Dispersion (absolue)	écart-type
Dispersion (relative)	coefficient de variation

Nous n'avons pas indiqué la variance $V = \sigma^2$ car notre expérience montre que certain(e)s ne se rendent pas compte que l'unité utilisée par la variance n'est pas la même que pour la variable. Ainsi, chaque année, nous voyons des intervalles $[m - V, m + V]$ mélangeant allégrement cm et cm^2 par exemple. A ces paramètres, on adjoint souvent, en informatique, à titre de vérification, les indicateurs *minimum* et *maximum*.

Lors d'une analyse où de nombreuses variables sont présentes, il est fréquent de se poser la question de l'ordre d'affichage des résultats. Les logiciels les moins intelligents se contentent de présenter les résultats variable par variable, dans l'ordre historique, c'est à dire dans l'ordre où elles ont été entrées dans le questionnaire, l'enquête ou le dossier. Nous préférons un ordre qui fait tout de suite ressortir les variables significatives. Pour les variables quantitatives, on viendra donc effectuer un premier affichage par ordre décroissant de coefficient de variation σ/m puis, dans certains cas (une même variable quantitative ventilée, variables de même ordre de grandeur...) un deuxième affichage par ordre décroissant de moyenne. Si les variables sont très hétérogènes ou incomparables entre-elles (comme taille, poids, age...) cela pourra, à défaut de permettre la comparaison, donner une indication sur un effet de taille possible.

Pour l'analyse conjointe, pour n variables, il a C_n^2 à couples de colonnes à traiter, soit $n.(n-1)/2$ valeurs à produire, nommés coefficients de corrélation linéaire, notés $\rho_{i,j}$ pour les variables Q_i et Q_j ou $\rho_{X,Y}$ pour les variables X et Y.

La corrélation linéaire entre deux variables quantitatives est la même que celle introduite pour les variables aléatoires. On calcule donc systématiquement tous les coefficients de corrélation linéaire entre variables quantitatives regroupés dans la *matrice de corrélation.* Il y a mathématiquement corrélation si $|\rho|$ est égal à 1. En pratique,[1] on dira qu'il y a corrélation pour $|\rho| > 0.9$. Il peut être intéressant (pour des phénomènes locaux) d'aller voir un peu plus loin. On viendra donc aussi fournir les couples (X,Y) de variables quantitatives telles que $|\rho| > 0.6$. Dans la mesure où ce sont les ordinateurs qui calculent, trient, affichent, on n'hésitera pas à consulter ces indications. Si X et Y sont en liaison linéaire, la relation de dépendance linéaire $Y = a.X + b$ est au mieux vérifiée pour

[C1] $a = \rho(X,Y).\sigma(Y)/\sigma(X)$
$b = m(Y) - a.m(X)$

Mais on peut aussi utiliser pour les formules :

[C2] $a = (m(X).m(Y) - m(XY))/d$
$b = (m(X).m(XY) - m(Y)mt(X,2))/d$
où $d = m(X)^2 - mt(X,2)$

Rappelons que la corrélation linéaire correspond à un cas particulier de *régression* où on cherche une relation ici sous forme d'une *droite* entre des variables. La méthode la plus simple ici pour calculer a et b consiste à minimiser l'erreur au sens des moindres carrés, à savoir $\Sigma(Y_i - (a.X_i + b))^2$.

On se méfiera de la symétrie mathématique de la relation de corrélation. Elle ne doit en aucun cas être confondue avec la notion de causalité. S'il est vrai que plus X et Y sont liés (linéairement), plus $|\rho(X,Y)|$ augmente, cela n'indique pas pour autant que X implique Y ou que Y implique X. Que le nombre de couches-culotte achetées chaque année augmente dans la même proportion que le nombre de voitures le montre clairement. Il peut y avoir une troisième variable (par exemple la "croissance économique") qui est cause des deux autres. Attention aussi à la transitivité faible de la relation de corrélation linéaire (parfois surprenante, dans les listings, pour les débutants).

1. c'est en fait beaucoup plus compliqué que cela, mais à un niveau élementaire, un seuil constant comme 0.9 est suffisant.

3.3 Traitement des Variables qualitatives

Pour une variable qualitative, on effectue (et le nom peut sembler curieux pour l'instant) comme analyse séparée un *tri à plat*. Il s'agit en fait de comptages, que l'on double souvent de fréquences (ou pourcentages). Les notations classiques pour l'analyse d'une variable qualitative sont très simples à mémoriser : le tri à plat d'une telle variable avec m modalités numérotées 1,2... m utilise les codes c_j, j de 1 à m pour dénombrer l'effectif absolu n_j (ou "nombre de fois") de la modalité numéro j qu'on présente via son label (ou "libellé" l_j). Le total $Nt = \sum n_j$ permet de calculer les fréquences (relatives) $f_j = n_j/Nt$ et les proportions (ou pourcentages) $p_j = 100 * fj$.

Il ne faut pas donner seulement les pourcentages sans préciser le nombre total d'individus car sinon on masque la représentatitivé de l'échantillon vis à vis de la population. En d'autres termes, à partir des n_j on peut calculer les f_j mais la réciproque est fausse : à partir des f_j on ne peut pas calculer les n_j (mais on le peut si on connait les f_j **et** la somme Nt des n_j).

Au niveau d'un affichage *intelligent* pour les variables qualitatives, on viendra tout d'abord ordonner chaque chaque liste de modalités d'une même variable par ordre décroissant d'importance (avec certainement le pourcentage entre parenthèses) puis on viendra classer ces variables par ordre de plus forte modalité. Là encore, ceci n'aura qu'un sens si ces variables ne sont pas trop différentes les unes des autres : il est clair qu'une variable comme "profession", "CSP" avec 15 rubriques ne sera pas aussi remplie par modalité qu'une variable comme "sexe" avec au plus 3 modalités.

Pour des variables qualitatives dites ordinales (c'est à dire ordonnnées), la notion de corrélation peut avoir un sens (corrélation au sens de *Kendall*, de *Spearman*), mais nous ne les traiterons pas ici. La généralisation des comptages donne lieu à des tri croisés. Le terme tri à plat s'explique alors par le fait que les totaux en lignes, en colonnes des tris croisés (donc à deux variables) donnent précisément les comptages à une seule variable. Un tri à plat s'obtient donc en "aplatissant" un tri croisé.

On peut raffiner les tris croisés en ne donnant pas seulement les comptages mais aussi les différents pourcentages (par rapport au total général, par rapport au total de la ligne, par rapport au total de la colonne...).

Il n'est pas bon en général de donner tous les tris croisés car certains sont biaisés. Ainsi, croiser *sexe* et $profession$ n'est pas très juste sous l'angle de l'équirépartition : certaines professions emploient principalement des femmes, d'autres sont plutôt réservées (à juste titre?) aux hommes... De même en ce qui concerne le croisement entre niveau d'étude et classe d'age. Il est difficile de s'attendre à trouver des personnes de moins de 15 ans au niveau troisième cycle des universités... On vient donc souvent choisir – et c'est un arbitraire – quelques tris croisés, espérant qu'ils seront significatifs, révélateurs.

Si l'ordre d'affichage des corrélations linéaires est facile à trouver (on utilise l'ordre décroissant des $\rho(X,Y)$ en valeur absolue), il n' y a pas pour l'instant de "bon choix" pour chacun des tableaux croisés. Pourtant, à la réflexion et à la lumière de la notion de calcul de χ^2, le lecteur pourrait déjà en proposer un ou deux, pas très immédiats. Nous laissons ces ordres à faire à titre d'exercice.

On doublera très souvent ces analyses chiffrées par des graphiques, que ce soit des courbes de valeurs, histogrammes de données ou de fréquences. Attention toutefois pour les variables quantitatives à ce qu'un découpage en classes n'est pas chose aisée. Le choix du nombre de classes, des bornes de classes n'est pas ni simple ni systématique (même si certains logiciels les tiennent pour "évidentes").

3.4 Traitement des Variables textuelles

L'analyse séparée d'une variables textuelles construit les dictionnaires des "mots", par ordre alphabétique et par odre d'occurrence (avec comptages et pourcentages). La linguistique nous apprend qu'il faut distinguer comme mots les formes graphiques (chaines de caractères telles quelles) des lemmes (ou "racines" comme le lemme *chant* pour les formes *chants*, *chanteur*, *chanteuses*...), qu'il faut repérer les *happax* (mots qui n'apparaissent qu'une seule fois dans un texte). L'analyse conjointe s'intéresse aux concordances c'est à dire aux positionnements relatifs des mots dans les phrases. Ainsi le mot "étrangers" dans la phrase "vive les étrangers!" et et dans la phrase "expulsons les étrangers" sont des environnements très différents du même mot "étrangers").

Signalons une autre difficulté : la transcription de l'oral est plus "fidèle" dans une orthographe aménagée que dans une écriture coventionnelle, mais oblige à des codages personnalisés.

3.5 Rédaction et Interprétation

La rédaction, l'interprétation des moyennes, comptages... obtenus est une phase obligatoire du traitement statistique. Elle coute parce qu'elle est difficile et engage son auteur. L'équation

$$\mathsf{Rédaction} = \mathsf{Traitements} + \mathsf{Interprétations}$$

doit rappeler qu'il faut d'abord calculer et ensuite interpréter, qu'une partie sans l'autre aboutit à une analyse statistique incomplète.[2] Rédiger, c'est calculer et dire qu'on a calculé, c'est dire qu'on a regardé les résultats, qu'on essaie de prendre partie, qu'on n'est pas une machine qui produit des chiffres et des chiffres encore... Rédiger un commentaire sur une moyenne, l'écart-type et le coefficient de variation, c'est interpréter des différences, des ressemblances. Une consommation moyenne de 600 litres de vins par an c'est beacuoup pour un particulier, c'est peut être peu pour une cantine de 30 employés, c'est peut-être beaucoup pour une autre cantine avec aussi 30 employés. Rédiger est un art, bien rédiger se fait avec le commanditaire de l'analyse, sous la validation d'experts du domaine, en accord avec les utilisateurs, les professionnels et habitués du chap investigué.

Sugnalons qu'en n particulier, lors de calculs d'effectifs, on se méfierade l'équirépartition, base de nombreux sondages qui peut être criticable ; ainsi, pour un dossier comme ELF, la disproportion apparente entre hommes et femmes (1 homme intérrogé pour 2 femmes interrogées) est justifiée par le thème de l'enquête (la Féminisation du nom des métiers).

La fausse précision scientifique assurée par de nombreuses décimales doit souvent être remplacée par un arrondi accompagné d'un commentaire d'imprécision. Ainsi "environ un tiers..." sera bien mieux perçu et plus facile à se rappeler, à transmettre que 32.7869 %. Il faut aussi rappeler les modes de calcul utilisés. Lors d'une élection récente sur les questions européennes, on a affiché partout qu'il y avait une majorité de "oui à l'Europe" alors que les chiffres étaient 51 % de oui, 49 % de non et 12 % d'abstention. Est-ce vraiment une majorité? Peut-on employer les termes de victoire, d'affirmation forte et collective avec une si faible différence de réponses?[3]

2. C'est sans doute ce qui explique pourquoi les mathématiciens s'arrêtent dès qu'on dispose du fichier de données en disant que ce n'est plus de la statistique mais de l'informatique et ce qui explique aussi pourquoi les informaticiens s'arrêtent lorsque les calculs sont effectués, en disant que ce n'est plus de la statistique mais de la littérature...

3. et avec un total de 112 %!

C'est là tout un art que de "noyer le poisson " (et l'électeur avec) que de glisser des chiffres aux phrases. Nous ne saurions trop encourager le débutant à déjouer les pièges d'une fausse science statistique qui joue sur les mots pour masquer les problèmes. Ainsi, lors de débats, notamment télévisés, il est fréquent de voir des discussions sur le nombre exact de chômeurs, qui comptant en "données normalisées", qui comptant "après correction des données saisonnières". Il est évident qu'avec deux méthodes de comptages différentes, les résultats sont différents. Mais la réponse à la question "Combien y a-t-il de chômeurs", la réponse devrait être unanime : "Il y en a trop", que ce soit 3,2 millions avec une méthode ou 3,4 millions avec une autre.

3.6 Exemples de traitements

Reprenons notre dossier Vins, extrait du J.O. en date du 4 novembre 1987, suite à une demande de Statistique Mensuelle des Vins par la Direction Générale des Impots, qui donne les importations pour 18 catégories de vins (en ligne) dans 8 pays (en colonne) et dont les données sont :

	BELG	NEDE	RFA	ITAL	UK	SUIS	USA	CANA
CHMP	7069	3786	12578	8037	13556	9664	10386	206
MOS1	2436	586	2006	30	1217	471	997	51
MOS2	3066	290	10439	1413	7214	112	3788	330
ALSA	2422	1999	17183	57	1127	600	408	241
GIRO	22986	22183	21023	56	30025	6544	13114	3447
BOJO	17465	19840	72977	2364	39919	17327	17487	2346
BORG	3784	2339	4828	98	7885	3191	11791	1188
RHON	7950	10537	7552	24	8172	11691	1369	1798
ANJO	2587	600	2101	0	7582	143	872	131
AOCX	17200	22806	15979	50	20004	1279	4016	944
VDQS	1976	1029	1346	0	2258	212	1017	487
XXXX	38747	19151	191140	7992	101108	1029	26192	38503
PROV	1375	1150	2514	0	284	401	9	236
MUSC	2016	2908	1529	0	12891	18	716	653
RHOF	785	1648	1009	6	775	643	542	35
AOCF	160	246	135	8	1177	26	7	0
XXXF	24	1533	160	0	480	0	0	0
XXFF	2415	74	208	8	1705	12	36	47

Un habitué des Statistiques et de l'Analyse des Données verrait [4] à l'oeil nu, en lisant soigneusement ce tableau, les grandes lignes de son ASG [5].

Regardons ce que donne déjà l'analyse séparée. Un affichage mécanique, inintelligent, par ordre historique d'entrée des colonnes fournit les résulats

Champ	Nom du champ	Moyenne m	Ecart-Type s	s/m	Cdv	Min	Max
1	NUM	non num. ; son type est : N					
2	BELGIQUE	7470	9994	134	24	38747	38723
3	NEDER	6261	8228	131	74	22806	22732
4	RFA	20261	44616	220	135	191140	191005
5	ITALIE	1119	2511	224	0	8037	8037
6	UK	14299	23616	165	284	101108	100824
7	SUISSE	2965	4882	165	0	17327	17327
8	USA	5153	7337	142	0	26192	26192
9	CANADA	2814	8705	309	0	38503	38503

Puisque les colonnes traitent d'*une même variable quantitative* ventilée par catégorie de vin et par pays, effectuons un tri par ordre décroissant de moyenne :

Champ	Nom du champ	Moyenne m	Ecart-Type s	s/m	Cdv	Min	Max
1	NUM	non num. ; son type est : N					
4	RFA	20261	44616	220	135	191140	191005
6	UK	14299	23616	165	284	101108	100824
2	BELGIQUE	7470	9994	134	24	38747	38723
3	NEDER	6261	8228	131	74	22806	22732
8	USA	5153	7337	142	0	26192	26192
7	SUISSE	2965	4882	165	0	17327	17327
9	CANADA	2814	8705	309	0	38503	38503
5	ITALIE	1119	2511	224	0	8037	8037

4. ...et commencerait certainement par critiquer ou au moins demander des explications sur le choix des pays, la typologie utilisée pour définir les catégories de vins...

5. *Analyse Statistique Générale.*

On met alors tout de suite en évidence un *très gros importateur* `RFA`[6] et un *très faible importateur* `ITALIE`. Pour savoir comment varient les importations par pays, effectuons maitenant un deuxième tri, par ordre décroissant de coefficient de variation ce qui nous donne

Champ	Nom du champ	Moyenne m	Ecart-Type s	s/m	Cdv	Min	Max
9	CANADA	2814	8705	309	0	38503	38503
5	ITALIE	1119	2511	224	0	8037	8037
4	RFA	20261	44616	220	135	191140	191005
6	UK	14299	23616	165	284	101108	100824
7	SUISSE	2965	4882	165	0	17327	17327
8	USA	5153	7337	142	0	26192	26192
2	BELGIQUE	7470	9994	134	24	38747	38723
3	NEDER	6261	8228	131	74	22806	22732

On peut en déduire que `CANADA`, `ITALIE` et `RFA` sont des importateurs *sélectifs*, important surtout du vin de catégorie `XXXX` (mais aussi du `CHMP` pour `CANADA`).

On remarquera l'oeuvre du statisticien par rapport au travail de l'informaticien : en informatique on produit les résultats, en statistiques on les trie et on les commente.

Passons maintenant à l'analyse conjointe, c'est à dire l'analyse par couple de variables. La matrice des corrélations, fournie sous forme "triangulaire améliorée" pour des raisons pédagogique) est

	BELGI	NEDE	RFA	ITAL	UK	SUISSE	USA
NEDE	0.8702	1.0000					
RFA	0.8692	0.5818	1.0000				
ITAL	0.5856	0.2895	0.6998	1.0000			
UK	0.9416	0.6997	0.9693	0.6906	1.0000		
SUIS	0.3353	0.5177	0.1984	0.3098	0.2462	1.0000	
USA	0.8699	0.6799	0.8477	0.7172	0.8935	0.4681	1.0000
CANA	0.8143	0.4582	0.9476	0.6585	0.9256	-0.0246	0.7469

6. pour les jeunes qui lisent ce texte, il s'agit d'une partie ce qu'on nomme aujourd'hui l'Allemagne.

Et on peut réduire la liste des meilleures corrélations linéaires par ordre décroissant de $|\rho|$ à :

```
0.969 UK    RFA
0.948 CANAD RFA
0.942 UK    BELGI
0.926 CANAD UK
```

Les valeurs des formules linéaires pour les corrélation dites "sûres" sont alors

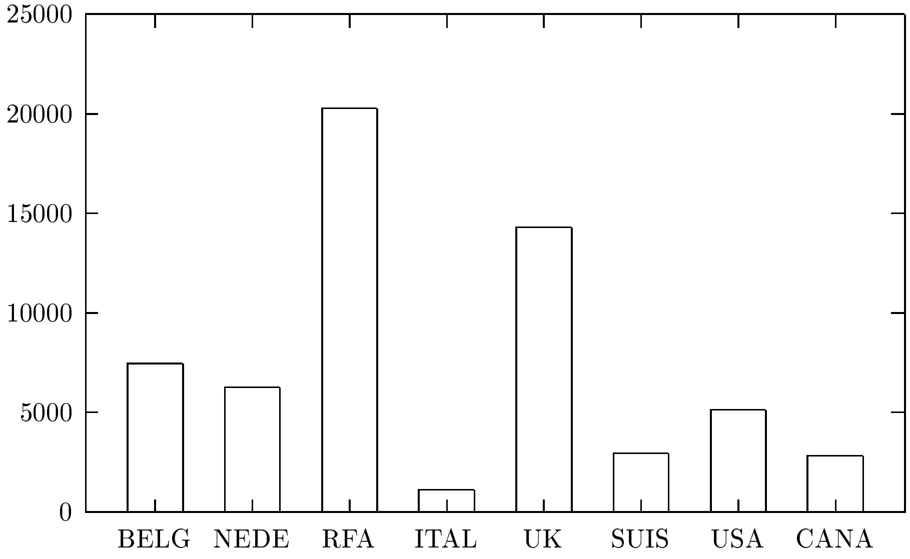
```
0.969  :  UK      =  0.513  *  RFA       +  3903.587
0.969  :  RFA     =  1.831  *  UK        -  5921.349
0.948  :  CANADA  =  0.185  *  RFA       -   932.386
0.948  :  RFA     =  4.857  *  CANADA    +  6596.411
0.942  :  UK      =  2.225  *  BELGIQ    -  2322.591
0.942  :  BELGIQ  =  0.398  *  UK        +  1772.722
0.926  :  CANADA  =  0.341  *  UK        -  2064.845
0.926  :  UK      =  2.511  *  CANADA    +  7233.245
```

L'interprétation de ces calculs par couples peut se résumer à ceci : quatre pays semblent se comporter de façon assez similaire (au sens du modèle linéaire), à savoir `UK`, `RFA`, `CANADA` et `BELGI`. Ainsi, on peut estimer que `RFA` importe[7] très grossièrement de 1,5 à 2 fois ce qu'importe `UK`.

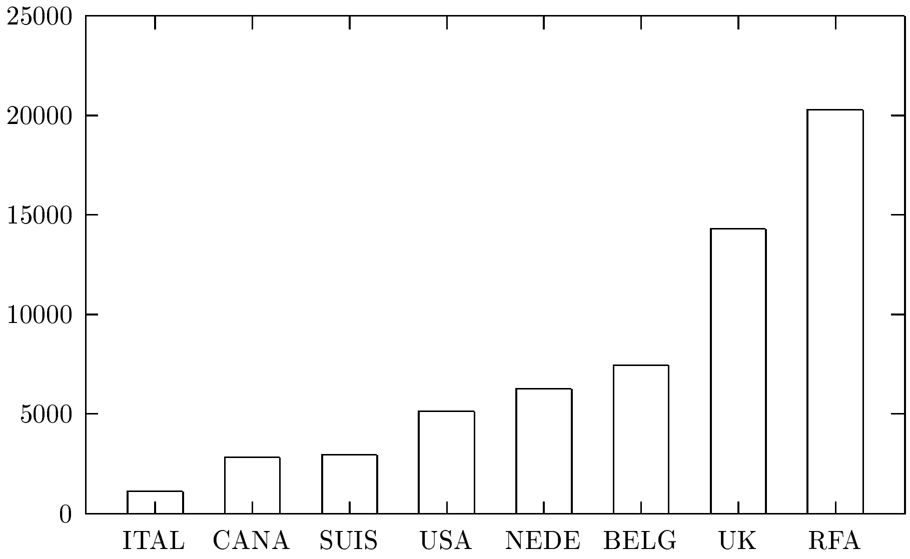
Pour valider, mieux faire ressortir nos commentaires, des graphiques seraient les bienvenus, comme celui des moyennes d'importation par pays, celui des droites de corrélation linéaires entre `UK` et `RFA`...

7. un sujet d'examen comportait un jour une question sur ce que `RFA` *consomme* et non pas ce que `RFA` *importe*, mais beaucoup d'étudiant(e)s malheureusement tombèrent dans ce piège linguistique qui invalide l'étude statistique...

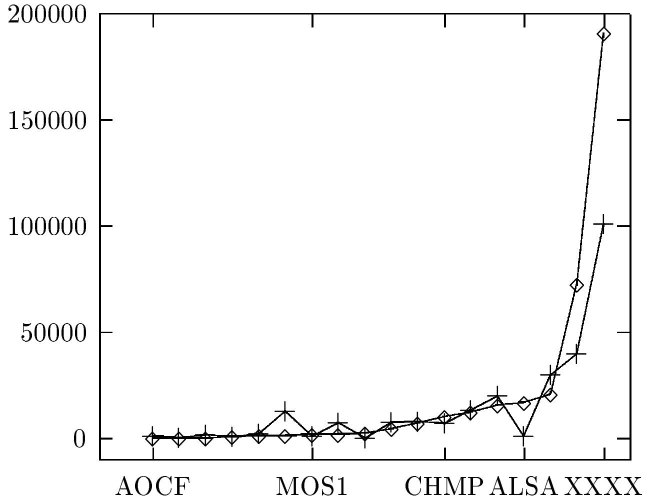
Pour le graphique des moyennes d'importation par pays, on a le choix entre utiliser l'ordre historique des pays



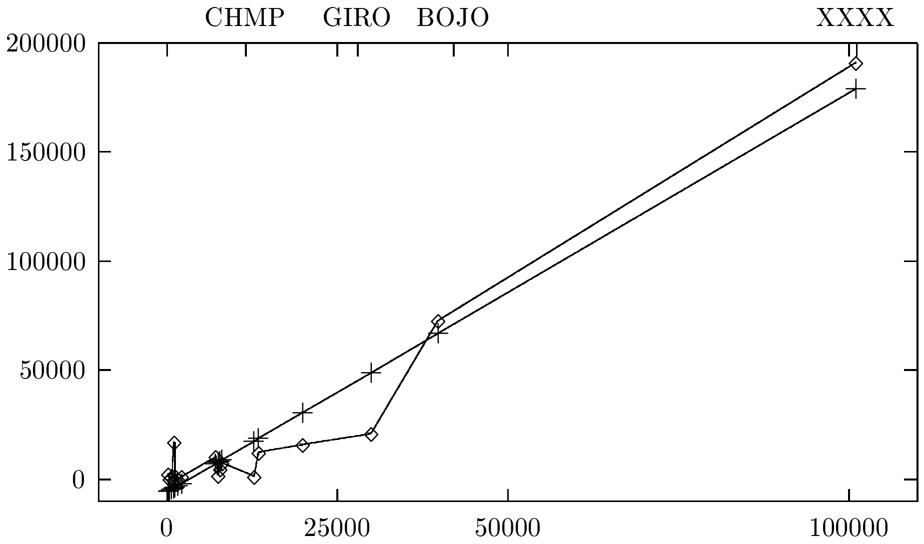
et l'ordre croissant de moyennes



Pour les droites de corrélation linéaires, nous ne présentons comme exemple que le couple (`UK`,`RFA`) avec les courbes comparées de `UK` (+) et `RFA` ($\diamond$)



la courbe $RFA = f(UK)$ avec les symboles $\diamond$ et la droite de régression $\widehat{RFA} = 1{,}831.UK - 5921.349$.



Pour un véritable exposé, il faudrait discuter ici de la couleur, des polices de caractères, intégrer les graphiques dans le discours etc.

Comme exemple de traitement pour les variables qualitatives, nous prendrons le dossier CETOP nommé ainsi par que nous avons mené cette étude avec et pour le *Centre d'Etudes et Techniques d'Optimisation des Processus*. Après l'enquête effectuée par publi-postage, nous avons obtenu 246 questionnaires remplis exploitables avec chacun une centaine de réponses cochées.[8]. Le questionnaire comportait 5 variables quantitatives, 45 variables qualitatives. Nous en extrairons seulement quelques-unes, particulièrement remarquables, à savoir le code `SEXE` de l'individu, le niveau d'études (noté ici `FORM`), le type d'études (`SPEC`), l'appartenance (`APPQ`) à un *Cercle de Qualité* et le contexte de l'entreprise (`CONT`).

L'affichage mécanique dans l'ordre des variables de départ et par ordre de code croissant donne ce magnifique affichage illisible

```
SEXE 0  16 (  6.50 %)  1 215 ( 87.40 %)  2  15 (  6.10 %)
FORM 0  17 (  6.91 %)  1  10 (  4.07 %)  2  14 (  5.69 %)
     3 205 ( 83.33 %)
SPEC 0  34 ( 13.82 %)  1 146 ( 59.35 %)  2   1 (  0.41 %)
     3  22 (  8.94 %)  4   6 (  2.44 %)  5  29 ( 11.79 %)
     6   8 (  3.25 %)
APPQ 0  20 (  8.13 %)  1 101 ( 41.06 %)  2 125 ( 50.81 %)
CONT 0  38 ( 15.45 %)  1  13 (  5.28 %)  2  54 ( 21.95 %)
     3 141 ( 57.32 %)
```

Alors que bien structuré, avec les bons libellés, on peut (presque) lire

```
NOM    Mode          Effectif           Mod.   Effectif
Sexe   HOMMES             215 ( 88 %)   nr             16 (  6 %)
Form   SUPERIEUR          205 ( 83 %)   nr             17 (  7 %)
       BAC                 14 (  6 %)   SECONDAIRE     10 (  4 %)
Spec   SCIENCES/TECHN.    146 ( 59 %)   nr             34 ( 14 %)
       QUALITE             29 ( 12 %)   GES            22 (  9 %)
Cont   INTERNAT           141 ( 57 %)   NATIONAL       54 ( 22 %)
       nr                  38 ( 15 %)   REGIONAL       13 (  5 %)
Appq   OUI                125 ( 51 %)   NON           101 ( 41 %)
```

où `nr` signifie *non-réponse*.

8. voici un exemple de mauvaise rédaction : ce qui est écrit laisse croire que tous les questionnaires n'ont pas forcment le même nombre de réponses cochées. En fait, la rédaction initiale était "avec chacun 122 réponses cochées utilisables" mais cela paraissait inutilement trop détaillée...

Cette petite analyse séparée met en évidence de nombreuses non-réponses, notamment au niveau de la variable SEXE. Essayons de comprendre ce que cela signifie... Les variables quantitatives du dossier montrent que le questionnaire moyen a été principalement rempli par des hommes de 40 ans en moyenne, avec un nombre moyen de 5 années d'études après le bac. Comment ces gens peuvent-ils ne pas savoir quel est leur sexe ? La réponse est ailleurs (sans vouloir copier le jeune *Mulder* des *X-files*). Le questionnaire, technique, long, était adressé aux décideurs, aux responsables d'entreprises. La notice incluse dans la page 1 du questionnaire indiquait qu'on pouvait le remplir à plusieurs. Quel code sexe faut-il mettre si 2 hommes et 3 femmes ont rempli le questionnaire ? Une telle maladresse dans le texte du questionnaire gâte visiblement la qualité des données, d'autres maladresses sont détectables au vu des autres tris à plat...

Donnons un seul tri croisé. Ce n'est pas la peine d'utiliser la variable SEXE car elle est très majoritairement dédiée aux hommes. Vouloir relier le contexte de l'entreprise au niveau d'étude parait un peu étonnant, essayons donc de croiser l'appartenance à un cercle qualité et le contexte. On obtient

```
                    NON      OUI    +  Total
nr           11      14       13    +     38
REGIONAL      1       5        7    +     13
NATIONAL      3      28       23    +     54
INTERNAT.     5      54       82    +    141
+++++++++++++++++++++++++++++++++++++++++
Total        20     101      125    +    246
```

Il semblerait, à la lecture du plus grand effectif dans le tableau, qu'on puisse en déduire que l'appartenance à un cercle qualité est liée au contexte international. Pour valider une telle hypothèse de dépedance, il faudrait réaliser un text de χ^2 mais ceci est une autre histoire...

Passons maintenant à un exemple de traitement de variable textuelle. Le texte traité est un compte-rendu d'un centre de placement pour enfants en difficulté. L'intrégralité du texte fait 338 lignes assez dense, soit un peu plus de 25 k de données. Nous reproduisons ici le début du texte, mettant [...] en lieu et place d'informations confidentielles :

```
par courrier du  [...] , notre service informe le juge des
enfants que  [...]   a été conduit  au centre éducatif  et
professionnel dès le     [...] ,  pour  un entretien  sans
engagement devant faciliter  la reprise du  contact, ainsi
que la négociation  de nouvelles possibilités  de travail.
rappelant  les  problèmes  sérieux  traversés  par   cette
institution, tel qu'évoqué dans le rapport d'investigation
et dans le cabinet du magistrat, le centre précise que  la
direction de l'établissement  reconnaît d'autant mieux  le
bien_fondé  de  nos  observations  que  la structure a été
encore  une  fois  éprouvée  par  des incidents d'une rare
violence,  dans  la  nuit  du   [...]   :  agression d'une
éducatrice  à   l'arme  blanche   ; intervention   de  la
gendarmerie  intra_muros  ;  violences,  débordements   et
insécurité des pensionnaires...
```

Le texte global comporte 4124 mots courants dont 1331 mots courants differents, dont le dictionnaire alphabétique commence par

```
  a                                    30   0.73
  abandonné                             1   0.02
  abord                                 1   0.02
  absence                               3   0.07
  académie                              1   0.02
  accepte                               2   0.05
  accepté                               2   0.05
  accident                              1   0.02
  accompagnement                        1   0.02
  accompagner                           1   0.02
  accompagné                            2   0.05
  accord                                1   0.02
...
  adolescent                           10   0.24
  adolescents                           1   0.02
```

Trié par occurences décroissantes, ce dictionnaire distingue les mots suivants (nous n'avons pas reproduit tous les articles et autres "petits mots ultra-fréquents impondérables"

```
   de                               212   5.14
...
   XXXX                              40   0.97
...
   ne                                36   0.87
   YYYY                              20   0.48
   mère                              19   0.46
   éducateur                         13   0.32
```

Le mot `XXXX` a été mis ici pour le nom de l'adolescent considéré et `YYYY` pour le nom du centre. Nous reproduisons maintenant une partie des concordances pour les mots `mère` et `adolescent`, c'est à dire son environnement à raisons de 3 mots avant et 4 mots après :

```
       reçu avec sa   mère    au centre, le 12
       pourquoi, la   mère    et l'enfant conviennent de
    c'est pas ;  sa   mère    ne répond pas à
               . La   mère    et l'enfant ne nous
       même avec ma   mère,   nous en avons jamais
 ...

       cas  où l'   adolescent    déciderait encore  de le
               L'   adolescent    nous déclare ne pas
               L'   adolescent    n'est  pas rentré  pendant
    de janvier l'   adolescent    est invité à nous
      le jeune      adolescent    avec un aplomb étonnant
   informe que l'   adolescent    est venu agresser verbalement
   charge de cet    adolescent    et de ses camarades.
```

Une de nos conclusions pour cette étude fut de renvoyer à une absence de "l"image du père", ceci pouvant expliquer cela, via une absence constatable du père (y compris de nos dictionnaires), mais aussi de ce qu'il représente (la loi, l'autorité...).

3.7 Autres Variables et autres Traitements

Un premier cas de variables numériques un peu moins classiques, et assez peu différentes des variables qualitatives est celui des **QM** c'est à dire des *variables multi-réponses.* Une variable qualitative serait par exemple

> "prenez-vous du café au petit déjeuner ? (oui/non)"

alors qu'une variable multi-réponses serait

> "que prenez-vous au petit déjeuner :"

- du café	(oui/non)	- du thé	(oui/non)
- du lait	(oui/non)	- du jeu d'orange	(oui/non)

Une façon simple de traiter une QM est de la décomposer en plusieurs QL. Il y a toutefois une petite difficulté : le nombre de réponses n'est pas toujours le même, ce qui empêche de faire des pourcentages avec un même total. On peut parfois imposer un nombre de réponses ("cocher une seule case"), donner une limite inférieure ou supérieure ("choisissez au plus 3 valeurs") mais cela ne résoud pas pour autant tous les problèmes.

Un deuxième type d'autres variables numériques, proches des variables quantitatives cette fois-ci est celui des **QP** ou *variables pourcentages.* Un exemple de variable pourcentage est fourni par un énoncé comme

> "comment répartissez-vous vos soiréees-télévision (en %) "

- sport	....
- films	....
- variétés	....
- autres	....

Comme pour les autres variables, il faut d'abord tester la cohérence des données. Il peut y avoir des non-réponses (faut-il leur mettre 0 % ?), des totaux incorrects (plus de 100 %, moins de 100 %). La question de savoir ce qu'il faut faire en cas de réponse incorrecte est délicate. Ignorer tous les pourcentages pour un total incorrect ou pour une non-réponse revient à changer le nombre total de personnes, c'est à dire la taille de l'échantillon. Si le nombre d'erreurs est important (ce qui est le cas pour des questionnaires en publi-postage), on peut arriver à des chutes d'effectifs très importants. Par contre, dans le cas de questions gérées *in situ* par un enquêteur, ce genre de problème ne se pose pratiquement pas.

Le traitement d'une QP ne pose pas *a priori* de difficultés. On effectue la moyenne de chaque pourcentage qu'on peut ensuite classer par ordre décroissant. Pour comparer plusieurs QP, on aura intérêt à indiquer clairement le nombre de réponses mises en jeu par variable. On peut aussi faire l'écart-type des réponses pour affiner la vision un peut trop synthétique fournie par les moyennes. Mais il est bon d'avoir une idée de ce que représente 1 %, 10 % car souvent il est difficile d'évaluer précisément de telles quantités.

Un pourcentage peut être remplacé par un *score*, ou un *anti-score*. Il n'y a alors plus de total imposé. Un exemple de question possible pour les scores est

"indiquez vos préferences en littérature par une note "
"de 1 à 5 (5 est la meilleure note)"

- romans	note :
- monographies	note :
- bandes dessinées	note :
- autres	note :

Pour des scores, il est en général conseillé d'indiquer le total général des réponses, de faire une étude quantitative des scores par personne de façon à connaitre un profil moyen de réponse (chaque modalité peut être fortement cotée, faiblement cotée...).

Une façon de normaliser les scores est d'utiliser un troisième type de variables, les **QH** ou *variables hiérarchiques*. On les nomme aussi *rangs*. On demande alors au questionné(e) de classer, par ordre de préférence ou d'anti-préférence des modalités[9] :

"classer de 1 à 3 ces chateaux sur la Loire"
"(1 est le meilleur)"

- Angers
- Chenonceaux
- Saumur

Même pour un petit nombre de modalités, des problèmes de cohérence peuvent surgir. Certain(e)s n'arrivent pas à se décider et veulent mettre des *ex-aequo*, d'autres se trompent finalement sur le choix du dernier, etc. Une fois ce genre d'erreurs traitées, on peut se ramener à un calcul de moyenne.

9. où est le piège?

Une variable **QF** ou *Question Floue* est une variable à modalité pour laquelle on attribue une valeur probabiliste (nombre réel entre 0 et 1, dont la somme fait 1) à chaque modalité. Une **QF** peut par exemple résulter d'une **QT**. Ainsi la variable AGE peut se découper en "jeune" (moins de 30 ans) et "vieux (plus de 30 ans). Ce codage brutal est adouci par un codage flou. Ainsi pour une personne de 35 ans, la pondération jeune=0.8 et vieux= 0.2 indique que l'individu est plutot jeune mais un peu vieux quand même. Le découpage en classe classique aurait donne ici jeune=0 et vieux=1.

On peut aussi calculer des comptages pour un dernier type de variables que nous nommerons **QE** ou *variable d'énonciation*. Ce genre de variables intervient dans le cas de question ouverte, c'est à dire où les réponses ne sont pas imposées, contrairement à toutes les variables vues jusqu'ici, qui étaient des questions fermées. Un modèle du genre est par exemple :

> "citez des marques de voitures japonaises"

Il faut d'abord vérifier l'exactitude des réponses avant de comptabiliser les bonnes réponses. On obtient alors des variables proches des QT à ceci près qu'elles ne prennent que des valeurs entières. Ces variables sont alors assez proches des **Questions ouvertes**, plus délicates à mettre en oeuvre et qu'on peut inclure dans les *variables textuelles*. Un exemple type de variable qualitative, de variable quantitative et de variable textuelle (et on réfléchira à la quantité, à la qualité et à la gradualité de l'information fournie par les réponses) peut être :

> "avez-vous une activité salariée en ce moment ?"
> - oui
> - non
> - autre
>
> "à votre avis combien de cadres sont salariés en France ?"
>
> "donnez en quelques lignes, selon vous, les raisons du "
> "chomage des cadres en France"

Apportons au passage quelques compléments à ce que nous avions dit sur les variables textuelles. Le point d'entrée est trés général : une QX peut être une réponse à une question ouverte, un texte complet, un récit, un compte-rendu d'expertise [10] ...

10. Il n'y a jamais obligation de répondre à un questionnaire. D'autre part certaines questions peuvent être gênantes, ou peuvent servir à des fins autres que statistiques... C'est pourquoi s'il faut toujours prévoir des non-réponses (même pour une simple question en oui/non), il faut aussi accorder une importance à l'absence de réponse pour une QX.

Traiter une QX est d'un tout autre ordre que traiter une QT ou une QL. Nous avions indiqué qu'il y a plusieurs niveaux d'analyses possibles : le niveaux lexical (ou syntaxique) et le niveau lemmatal (ou sémantique) suivant que l'on décide de traiter les "mots" ou les "concepts". Dans le premier cas, on parle en fait de forme graphique ou de chaine de caractères, dans le second de lemme ; notre usage est d'employer le terme "élément de texte" pour regrouper les deux vocables forme graphique et lemme. En mode non lemmatisé, "chanteur", "chanteuse" et "chanteuses" comptent pour trois formes différentes. Si on lemmatise, ce seront trois occurrences du lemme "chanteur". Pour travailler avec les lemmes, il y a un travail préparatoire qui consiste au moins à mettre tous les substantifs au masculin singulier, à remplacer les formes par les lemmes, à passer tous les verbes à l'infinitif...

Une fois le texte préparé les premiers calculs en *lexicomètrie* ou en *statistique lexicale* consistent en des comptages d'occurences (de formes ou de lemmes suivant le mode de traitement choisi). Les résultats sont présentés dans des *lexiques* ou *dictionnaires* qui sont des tableaux à deux colonnes, une pour les éléments de texte, l'autre pour leur nombre d'occurrences (nous avons déjà donné un extrait de tels dictionnaires dans la section 3.6).

La difficulté lors de l'exploitation des résultats d'analyses de textes[11] est que l'on doit apprécier les comptages, pourcentages... L'être humain est, notamment mais pas seulement, plus dans la diversité (la mutiplicité homogène, l'unicité hétérogène) que dans la stricte rigueur de la norme. Ainsi comparer une largeur à une cote même avec une tolérance ne relève pas de la même appréciation que la comparaison de l'expression d'un sentiment vis à vis d'une situation.

C'est aussi pourquoi on s'intéresse autant aux termes les plus fréquents (mais certains sont prévisibles[12]) qu'aux *happax* qui sont les termes qui n'apparaissent qu'une seule fois. Le nombre de mots différents et le rapport du nombre de mots différents sur le nombre de mots en tout fournissent alors des éléments d'appréciation de la richesse du vocabulaire. Après les comptages élémentaires, on s'intéresse aux *environnements* c'est à dire à la localisation des éléments de texte : après avoir effectué un choix (arbitraire) de mots significatifs, qui sont souvent les mots les plus fréquents, on vient extraire de la QX les portions de phrase qui entourent ce mot, avec un nombre fixe de mots avant et un nombre fixe de mots après.

11. Les statisticiens les nomment "analyses (de données) textuelles", les linguistes "analyses de contenu", les informaticiens "*data mining*", mais les buts de ces analyses ne sont pas rigoureusement les mêmes

12. Lesquels ?

Il reste ensuite à essayer de décrire ce que cela semble induire en fonction des environnements, de ce que cela éveille en nous... Une analyse de données textuelles demande donc (même si elle ne le présuppose pas, car sinon ce serait la réserver à une *élite* de gens doués ou prédisposés) de la sensibilité, une ouverture d'esprit, une culture, un esprit d'analyse et de synthèse qu'un(e) débutante peut acquérir grâce au travail en groupe, à la confrontation mutuelle d'appréciations. Ce genre d'analyse est donc souvent effectué de concert avec une étude sociologique, socio-linguistique, psycho-historique etc.

Une sélection de mots permet aussi d'effectuer une analyse des données (AFC) sur le tableau qui croise unités de texte et occurrences des mots choisis. Les plans factoriels fournissent alors une vision simpliste mais visuelle des liaisons entre éléments de texte et unités de texte... La statistique lexicale est plus délicate à mettre en oeuvre que la statistique numérique traditionnelle car elle suppose des connaissances sur la langue, sur les langues. En particulier, les comptages de base ne doivent pas être lus de la meme façon d'un pays à l'autre. Ainsi, en italien les pronoms personnels comme je, tu, il... sont beaucoup moins présents qu'en français car la personne est indiquée par la terminaison du verbe...

3.8 Autres études statistiques

Deux questions toutes naturelles qui se posent lors de l'analyse de plusieurs variables qualitatives sont celles de l'égalité de deux répartitions et l'adéquation (ou approximation) de la répartition de ces variables par des lois usuelles. Si la comparaison mathématique (stricte égalité terme à terme) ne peut pas être retenue pour des raisons pratiques, il faut chercher d'autres indicateurs numériques de ressemblance.

On utilise en général un seul indicateur global, ce qui oblige souvent à utiliser une expression à base de sommes de différences. Le calcul classique à effectuer dans ce cas est nommé χ^2 qui met en jeu des différences pondérées (relatives) d'où une importance moindre pour une même différence en cas d'effectifs importants.

Ce peut être un χ^2 d'adéquation pour indiquer à quel point une série d'effectifs *observés* correspond à une série d'effectifs *théoriques*, un χ^2 d'indépendance sur un tableau de contingence pour tester la dépendance entre des effectifs liés à deux variables qualtitatives... Plus généralement, la notion de *test* vient valider ou invalider une série d'hypothèses, Un test requiert une méthodologie stricte, avec le choix des hypothèses à tester, la statistique (ou loi de probabilité) utilisée pour mesurer l'écart entre la valeur testée et la valeur estimée, la règle de décision à utiliser, le choix du seuil critique[13] le calcul du niveau de signification de la statistique observée. Nous renvoyons pour cela aux ouvrages de statistique classique.

Pour être complet, il faut aussi citer comme analyses statistiques classiques

- l'étude des *Séries Chronologiques* qui essaie de décomposer des données liées au temps en une tendance globale et des composantes saisonnières,
- l'*Analyse de Variance* qui sépare la variation totale d'un ensemble de données en composantes relatives à des modalités (ou des sous-populations), pour tester la [non-]égalité des moyennes "par blocs", et les *Plans d'Expérience*, qui comparent les variances réelles des coefficients d'un modèle aux variances des mêmes coefficients si les facteurs contrôlés sont sans effet,
- les techniques de *Régression* (linéaire, polynomiale, logistique, multiple,...) qui expliquent[14] une variable quantitative à partir d'une ou plusieurs variables quantitatives via un modèle d'équation(s) donné à l'avance.

13. il faudrait ici définir la notion de risque α de première espèce (qui est la probabilité de rejeter l'hypothèse alors qu'elle est vraie), la notion de risque $\beta(\theta)$ de deuxième espèce (qui est la probabilité d'accepter l'hypothèse pour la valeur θ alors qu'elle est ...)

14. ce terme est à prendre au sens mathématique, et à notre sens, c'est donc quelque part une escroquerie, car cela signifie seulement donner une équation.

Chapitre 4.

Logiciels statistiques

4.1 Informatique, Programmation statistique

De nombreux logiciels de statistiques sont disponibles sur le marché qui contiennent tout ce dont un statisticien a besoin. On peut notamment citer *SAS*, *StatGraphics*, *Statistica*, *StatLab*, *StatItcf*, *Bmdp*, *Minitab*, *Spss*... Souvent les statistiques descriptives y cotoient l'Analyse des Données et les autres traitements statistiques que nous avons évoqués au chapitre 3, comme la Régression, l'Analyse des Séries Temporelles... Si ces logiciels sont souvent complets, ils n'en sont pas moins faciles à utiliser. Le principe de mise en oeuvre est souvent le même : après avoir chargé un ou plusieurs fichiers de données, on lance une option d'un menu ou d'un sous-menu qui demande éventuellement sur quelle sélection des données il faut effectuer le traitement.

D'autres logiciels classiques de traitement de données, de base de données ou des tableurs, des gestionnaires de feuilles de calculs sont largement suffisant pour des statistiques élémentaires. Ainsi *Excel* et *Dbase* contiennent sufisamment de fonctions sur lignes et colonnes d'une base de données pour fournir très rapidement moyennes, variances, écart-types, tris à plat, tris-croisés... quand ces opérations ne sont pas déjà intégrées au logiciel suivant la version utilisée ! De même les grands logiciels de gestion de bases de données comme *Oracle*, le langage *SQL* permettent de manipuler les données, d'effectuer les calculs, soit par l'appel d'instructions de base, soit en reprogrammant les calculs désirés.

Il faut en général se méfier des logiciels spécialisés en statistiques, non sur la validité ou la qualité des résultats, mais sur la présentation des options, la mise en forme des valeurs chiffrées.

L'utilisateur débutant peut être noyé sous l'avalanche des paramètres, des options, des sorties chiffrées avec leurs indices complémentaires, leurs coefficient de déviation d'erreurs, de déviation-standard et autres R^2, γ_3 etc. Un bon logiciel de statistiques devrait être capable de faire la différence entre 3, valeur d'une QT et 3, code pour une QL. Ce n'est malheureusement souvent pas le cas, et de nombreux débutants ont ainsi pu calculer des moyennes de numéro de téléphone ou de code-sexe, effectuer des tris à plats de taille ou de poids sans comprendre où était leur erreur. Il faut donc se méfier des opérations automatiques effectuées sur l'ensemble des fichiers, surveiller les mises à jour automatiques, les calculs globaux...

De plus, avec ce que nous avons dit sur la volonté de décrire les variables, leur origine et leur signification, il est clair que chaque colonne de valeurs devrait pouvoir être renseignée, que les unités des QT devraient apparaitre au niveau des résultats, comme les libellés en clair des codes de modalités pour les QL. Ce n'est malheureusement pas souvent le cas. Pire même, certains logiciels numérotent les variables, les modalités (il est on ne peut plus illisible d'indiquer que "la moyenne de la variable Q017 est 25" alors que "la moyenne des ages est 25" est directement accessible). Il faut alors apprendre (quand c'est possible) à sauvegarder les résultats pour les reprendre sous éditeur de texte et remplacer ce qu'il faut...

Si l'on doit soi-même programmer des statistiques, on prendra soin à bien détailler ces renseignements. A titre d'exemple, prenons l'analyse simple de variables QT. Calculer la moyenne m, la variance v et l'écart-type s du tableau X contenant X valeurs notées $X[1]$,$X[2]$...$X[n]$ peut se faire facilement en traduisant dans le langage approprié l'algorithme de la page suivante.

```
MODULE AnaQt
 Globales   X        /* tableau des données  */
            n        /* nombre de valeurs    */
 Locales    k        /* indice de boucle     */
            sv       /* somme des valeurs    */
            sc       /* somme des carrés     */
 Fonction   externe :  rac  /* racine carrée  */
DEBUT DU MODULE StatDesc
  sv ← 0
  sc ← 0
  pour k de1a n
    sv ← sv + X[k]
    sc ← sv + X[k]*X[k]
  finpour k de1a n
  m ← sv/n
  v ← (sc/n) - m*m
  s ← rac(v)
FIN DU MODULE AnaQt
```

Mais s'il faut utiliser plusieurs variables, ne traiter par l'algorithme précédent que les QT, les nommer, gérer les unités, trier par ordre décroissant de moyenne puis par coefficient de variation décroissant, on aura plutôt comme algorithme quelque chose comme

```
MODULE StatDescQT

  Globales nbcol, typvare etc.

DEBUT DU MODULE StatDescQT
pour ic de1a nbcol
     tv ← typvar[ic]
     si tv = 'T' alors
        calcQt(moycol[ic], ectcol[ic], cdvcol[ic],
               mincol[ic], maxcol[ic], cdvcol[ic] )
     finsi
finpour
tridec(nomcol, ordrecol)
afficheQt(ordrecol, nomcol, moycol, ectcol, cdvcol,
          unitecol, mincol, maxcol)
FIN DU MODULE StatDescQT
```

Certains calculs mathématiques sont plus simples qu'il n'y parait. Ainsi le calcul des p premières valeurs de la loi de Poisson sachant le paramètre λ et le nombre de termes à afficher se fait simplement par

```
* Algorithme de la loi de Poisson

MODULE Poisson

 Paramètres lambda
            nbterm

 Locales    k        /* indice de boucle     */
            lk       /* puissance courante   */
            p        /* probabilite courante */
            fk       /* factorielle courante */

 Fonctions  externes exp        /* exponentielle            */
                     format     /* affichage avec décimales */

DEBUT DU MODULE Poisson

  écrire " Loi de Poisson "
  écrire "     paramètre lambda : " , lambda
  écrire "     nombre de termes : " , nbterm
  el ← exp(-lambda)
  k  ← 0
  lk ← 1
  p  ← el
  fk ← 1
  écrire " xi  puissance factorielle probabilité "
  écrire k,lk,fk,format(p*1.0000,10,4)
  tant que k <= n
     k  ← k+1
     fk ← k*fk        { factorielle }
     lk ← lk*lambda   { puissance   }
     p  ← p*lambda/k
     écrire k,lk,fk,format(p*1.0000,10,4)
  fin tant que

FIN DU MODULE Poisson
```

De même, l'affichage des valeurs de la loi binomiale, une fois les paramètres n et p donnés est réalisé par

```
* Algorithme de la loi Binomiale
écrire " Loi Binomiale B(n,p) "
q  ←  1-p
pk ←  1
qk ←  1
* calcul de q^n
k  ←  0
tant que k < n
   qk ←  q*qk
   k  ←  k+1
fin tant que
écrire " xi    c(n,k)   p^k   (1-p)^k   probabilité"
k   ←  0
cnk ←  1
écrire k,cnk,pk,qk,cnk*pk*qk
k ←  0
tant que k < n
   k   ←  k+1
   kk  ←  k
   pk  ←  p*pk
   qk  ←  qk/q
   cnk ← coefbin(n,kk)
   écrire k,cnk,pk,qk,cnk*pk*qk
fin tant que
```

où coefbin(x,y) désigne le sous-programme de calcul du coefficient du binome de Newton C_x^y.

Plus compliquées, par contre peuvent être certaines formules qui redonnent les tables et qui requièrent de fortes connaissances en Analyse Numérique. Par exemple, la fonction de répartition de la loi normale centrée réduite U est donnée par

$$f(u) = P(U < u) \simeq 1 - g(u).(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5)$$

Cette formule approxime la valeur exacte avec une erreur inférieure à 10^{-7} pour

$$
\begin{array}{lcl}
g(u) & = & e^{-u^2/2}/\sqrt{2\pi} \\
t & = & \texttt{1/(1+0.2316419*u)} \\
b_1 & = & \texttt{ 0.319381530} \\
b_2 & = & \texttt{-0.356563782} \\
b_3 & = & \texttt{ 1.781477937} \\
b_4 & = & \texttt{-1.821255978} \\
b_5 & = & \texttt{ 1.330274429}
\end{array}
$$

La traduction des algorithmes est plus ou moins facile selon les langages et les logiciels. Ainsi, sous *Dbase3*, il n'y a pas de boucle `pour`, il faut donc recourir à une boucle `tant que`. Par exemple, le programme Dbase qui correspond à l'algorithme pour la loi de Poisson, mis dans le fichier *Poisson.Prg* est

```
* Poisson.Prg
parameters lambda,nbterm
clear
set talk off
? " Loi de Poisson "
? "     paramètre lambda " , lambda
? "     nombre de termes " , nbterm
store exp(-lambda) to el
store  0  to k
store  1  to lk
store el  to p
store  1  to fk
? "       xi  puissance factorielle probabilité "
? k,lk,fk,str(p*1.0000,10,4)
do while k <= nbterm
   store k+1 to k
   store k*fk to fk
   store lk*lambda to lk
   store p*lambda/k to p
   ? k,lk,fk,str(p*1.0000,10,4)
enddo
```

Ce programme doit être appelé, si l'on veut les 12 premières valeurs de la loi de Poisson $\mathcal{P}(2.8)$ avec leurs probabilités respectives, par :

```
do poisson with 2.8,12
```

et il affiche alors

```
Loi de Poisson
    paramètre lambda          2.8
    nombre de termes          7
       xi       puissance factorielle    probabilité
        0          1.0000           1         0.0608
        1          2.8000           1         0.1703
        2          7.8400           2         0.2384
        3         21.9520           6         0.2225
        4         61.4656          24         0.1557
        5        172.1037         120         0.0872
        6        481.8903         720         0.0407
        7       1349.2929        5040         0.0163
        8       3778.0200       40320         0.0057
```

Comme pour tout traitement informatique, une analyse statistique et une analyse des données requiérent de l'organisation, un apprentissage... Pour être complet, ce travail doit se doubler d'une rédaction précise, d'une relecture critique et d'une présentation claire. C'est donc tout un art, comme la programmation, nécessaire ici aussi parfois pour automatiser les calculs, conversions, affichages et autres traitements...

4.2 Ce que devrait être un logiciel statistique

Les différentes phases d'une analyse statistique sont la définition des objectifs, la collecte des données, le traitement des données, l'interprétation des résultats et la rédaction de l'analyse (après avoir éventuellement bouclé sur la partie traitement et interprétation au vu des premiers résultats). Un logiciel statistique devrait permettre d'effectuer toutes ces phases.

De nombreux logiciels ne font aucune vérification sur le type et la nature des données. Ainsi *Excel*, comme tout tableur, laisse saisir des informations de nature différente, sans ordre de grandeur vérifié dans des cases (ou "cellules") contigues, enlevant toute homogénéité à la ligne, à la colonne utilisée. Utiliser un tableur pour des calculs statistiques oblige donc à beaucoup de rigueur et à beaucoup de méfiance.

La définition des variables, le choix de leur type, du nombre de décimales (en entrée, en sortie), est parfois inexistant, souvent restreint à des informations de nature informatique pour guider le stockage des données. Ainsi *Dbase* demande le nom et le type et la taille de la variable, proposant les types Caractère et Numérique, sans distinguer les valeurs numériques des variables qualitatives des valeurs numériques des variables quantitatives. Par contre, le bon point est que la taille des données filtre des fautes de frappe, que les masques de saisie y sont instantanés.

Pire, la plupart des logiciels ne font aucun lien quant à l'origine des données, le mode de recueil, la date de saisie etc. Notre notion de *fichier descriptif* ne semble pas exister dans la pratique, comme si seules les données comptaient, même si on ne sait pas à quoi elles correspondent !

Au niveau du traitement des données, les mêmes absences de garde-fous [1] se répétent : aucun logiciel ne semble demander "Avez-vous le droit de calculer la moyenne de ces valeurs ? " mais vient directement effectuer le calcul, même si les données sont des numéros de téléphone, des codes-sexe etc. Si on trouve parfois dans les fichiers d'aides des remarques sur le fait qu'il ne faut pas calculer seulement une moyenne mais aussi l'écart-type et le coefficient de variation correspondant, on ne voit pratiquement jamais d'aide à l'interprétation des résultats.

Ainsi une moyenne d'âge peut être : faible, très faible, [anormalement] élevée, peu différente de celle(s) observée(s) précédemment, sensiblement égale à ce qu'on en atttendait... De même, un effectif peut être important, majoritaire, totalement déséquilibré, surprenamment très élevé... La rédaction peut alors déboucher sur des termes comme équirépartition, équilibre, pondération, consensus, convergence ou au contraire relever l'hétérogénéité, la disproportion, la [forte] majorité, la disparité... S'il n'est pas possible qu'un logiciel choisisse de lui-même ces termes, il ne serait pourtant pas difficile d'afficher ces termes, en offrir des progressions graduées, des synonymes...

En ce qui concerne les fichiers, la rédaction, la mise en forme des documents, un logiciel d'analyse statistique devrait permettre de tout incorporer dans un seul document, ce qui faciliterait la reprise et la relance des calculs, voire leur automatisation. Un logiciel comme *StatView* propose des gabarits d'analyses qui sont des modèles d'enchainements de calculs non liés à un fichier de données mais à un modèle de fichier. De tels gabarits facilitent grandement le travail quand on doit par exemple effectuer les mêmes calculs statistiques chaque semaine sur un fichier différent mais avec une même structure...

1. on pourrait ajouter "de garde-folles" !

4.3 Le logiciel Dbase

Dbase, logiciel mythique des années 80, devint avec sa version *Dbase III+* dans les années 1985, un standard, une référence de format de fichiers. De simple logiciel de gestion de base de données, il est aujourd'hui devenu un logiciel de développement objet sous Windows connu sous le nom de *Visual Dbase* avec de nombreux concurrents et clone comme *Fox Base* [*Pro*], *Xbase*, après avoir été *Dbase IV*, *Dbase V Windows*... *Dbase* tire sa force de sa simplicité d'utilisation : il comporte un mode interactif dans lequel on peut taper des commandes à la main, les unes à la suite des autres et un mode programmé où on exécute la suite des mêmes actions mises dans un fichier.

Un autre grand intérêt de *Dbase* est la cohérence des donnée qu'il impose : une base de données comporte une structure et des données. La structure décrit les colonnes utilisées et impose un type à la colonne. Du coup, *Dbase* fournit avec ses fichiers `.dbf` un bon format d'échange de données. La saisie des données s'effectue alors directement par masque de saisie, diminuant les erreurs de frappe.

On ouvre une base de données sous *Dbase* avec la commande `USE`, on accède à la structure avec l'instruction `LIST STRUCTURE` lorsqu'elle est courte, par `DISPLAY STRUCTURE` lorsqu'elle est longue. La modification de la structure se fait en mode plein écran avec `MODIFY STRUCTURE`[2]. Les données peuvent être saisies au clavier avec `APPEND` ou lues dans d'autres fichiers par `APPEND FROM`. On affiche les données par `LIST` et par `DISPLAY`, on les modifie par `EDIT`, `BROWSE`, `REPLACE`...

Les calculs élémentaires se font sous *Dbase* avec `SUM`, `COUNT`, `AVERAGE` et `STORE`. Nous renvoyons à un cours plus de complet de *Dbase*, nous contentant de donner un exemple de script d'utilisation.

Le prompt de *Dbase* y est repéré par un point en début de ligne. La commande `SET ALTER TO` permet de définir un fichier comme sortie secondaire de l'affichage écran, la commande `SET ALTER TO` active ce fichier et `CLOSE ALTER` le ferme, permettant son édition ou son utilisation comme ici pour une démonstration. L'erreur sur la commande `COUNT` a été volontairement mise pour montrer que si les variables qualitatives sont déclarées en caractères, on ne peut pas les additionner, les moyenner..

2. pour aider l'utilisateur, les concepteurs de *Dbase* ont prévu qu'on puisse se contenter des quatre premières lettres des instructions. Ainsi `MODI STRU` équivaut à `MODIFY STRUCTURE`.

```
.set alter to dbase.dem
.set alter on

******************************************** exemple Dbase : QT

. use vins
. list stru
Structure de la base de données: C:vins.dbf
Nombre total d'enregistrements :      18
Date de la dernière mise à jour: 04/05/90
Champ  Nom champ   Type         Dim    Dec
    1  VIN         Caractère      4
    2  BELGIQUE    Numerique      7
    3  NEDERLAND   Numerique      7
    4  RFA         Numerique      7
    5  ITALIE      Numerique      7
    6  UK          Numerique      7
    7  SUISSE      Numerique      7
** Total **                      47

. list off
 VIN  BELGIQUE NEDERLAND     RFA  ITALIE      UK  SUISSE
 CHMP     7069      3786   12578    8037   13556    9664
 MOS1     2436       586    2006      30    1217     471
 MOS2     3066       290   10439    1413    7214     112
 ALSA     2422      1999   17183      57    1127     600
 GIRO    22986     22183   21023      56   30025    6544
 BOJO    17465     19840   72977    2364   39919   17327
 BORG     3784      2339    4828      98    7885    3191
 RHON     7950     10537    7552      24    8172   11691
 ANJO     2587       600    2101       0    7582     143
 AOCX    17200     22806   15979      50   20004    1279
 VDQS     1976      1029    1346       0    2258     212
 XXXX    38747     19151  191140    7992  101108    1029
 PROV     1375      1150    2514       0     284     401
 MUSC     2016      2908    1529       0   12891      18
 RHOF      785      1648    1009       6     775     643
 AOCF      160       246     135       8    1177      26
 XXXF       24      1533     160       0     480       0
 XXFF     2415        74     208       8    1705      12
```

```
. aver
     18 enregs moyenné(s)
BELGIQUE NEDERLAND     RFA  ITALIE      UK  SUISSE
    7470      6261   20262    1119   14299    2965

. count for rfa > 20000
      3 enregs
. disp  for rfa > 20000
Enreg. Nø  VIN  BELGIQUE NEDERLAND     RFA  ITALIE      UK  SUISSE
      5    GIRO    22986     22183   21023       56   30025    6544
      6    BOJO    17465     19840   72977     2364   39919   17327
     12    XXXX    38747     19151  191140     7992  101108    1029

. disp  vin for rfa > 20000
Enreg. Nø  vin
      5    GIRO
      6    BOJO
     12    XXXX

********************************************* exemple Dbase : QL

. use elf
. count to n
     99 enregs

. average sexe
Cette expression n'est pas numérique
             ?
average sexe

. count for sexe=0 to nbhom
Type de données incorrect
                  ?
count for sexe=0 to nbhom

. count for sexe="1" to nbhom
64 enregistrement(s)

. store nbhom*100/n to pcthom
             64.65

. ? " il y a " ,nbhom," hommes soit " ,str(pcthom,5,2)," % "
 il y a          64  hommes soit 64.65  %
```

```
. aver age to mage
     99 enregs moyenné(s)
 age
  36

. aver age*age to mcage
     99 enregs moyenné(s)
 age*age
    1589

. store mcage-mage*mage to varage
                   305

. store sqrt(varage) to ectage
        17.46

. close alter
```

4.4 Programmation statistique en Dbase

Les commandes précédentes peuvent être mises dans un fichier traditionnellement de type .prg, qu'on exécute par DO suivi du nom du fichier. *Dbase* version 3 possèdait en 1985 un éditeur de textes rudimentaires accessible via la commande modify command, dès 1994 *VisualDbase* dispose d'un éditeur très complet de textes et d'objets comme tout atelier logiciel de développement objet.

Le coeur du langage de dbase, avec `STORE` pour l'affectation, `DO WHILE` pour la boucle tant que (il n'y pas de boucle `POUR`), `IF` pour les tests simples et `DO CASE` pour les structures de choix a toujours été très simple à utiliser, même s'il est un peu lourd et verbeux comme *Cobol*. Les commandes `SET TALK`, `SET ECHO`, `SET STEP` avec leur paramètre `ON` ou `OFF` autorisent des traces et des déboggages rapides des programmes. La non déclaration des variables (et, ce qui va de pair, leur non typage explicite) en font un outil de développement agréable à utiliser. Comme exemple de développement, montrons comment analyser toutes les variables qualitatives d'un dossier stockées sous forme numérique avec tri des variables par effectif décroissant.

Pour un dossier comme *ELFB*, par exemple, avec les variables

SEXE	:	sexe de la personne (0 pour homme, 1 pour femme)
ETUD	:	niveau d'études (0 pour aucun diplome, 1 pour niveau sixième, (2 pour bepc, 3 pour bac, 4 pour post-bac)
ECIV	:	état civil (0 pour non-réponses, 1 célibataire,2 pour marié)

on veut aboutir à un affichage comme

```
nom   mode nb.    pct
SEXE    1   64  ( 64.65 %)   0   35    ( 35.35 %)
ECIV    1   56  ( 56.57 %)   2   34    ( 34.34 %)
ETUD    4   39  ( 39.39 %)   2   30    ( 30.30 %)
```

où nous n'avons fait figurer que les deux modalités de plus fort effectif.

Le programme utilise deux bases de données temporaires. La première est nommée asgqltmp.dbf et contient deux champs numériques NMOD et EFF dont le role est de trier le tri à plat d'une variable. La seconde, nommée asgql.dbf contient un champ caractère NOMVAR, deux champs numériques NBMOD et EFFTOT puis 20 champs numériques EFF01, EFF02...EFF20 pour stocker les effectifs (triés) de chaque modalité.

Le programme commence par demander le nom de la base, utilisant celui de la base ouverte par défaut et affiche éventuellement une aide à l'utilisation du programme. Une fois la vérification de l'existence du fichier effectuée, on commence enfin les calculs, mettant les résultats dans asgql.sor grâce à la commande `SET ALTER`. La boucle tant que qui correspond à l'instruction `do while fn > " "` où `store field(c) to fn` met le nom du champ numéro c dans la variable fn s'arrête lorsque que ce nom est vide, ce qui signifie qu'on a passé toutes les colonnes en revue. Au passage, l'instruction `if .not. type(fn)="N"` permet de ne traiter que les variables numériques, et les lignes

```
index on &fn to tmp
goto  top
store &fn to imin
goto  bottom
store &fn to imax
```

déterminent les valeurs minimales et maximales des numéros de modalité.

Avec les instructions

```
count for &fn=i to n
...
append blank
replace nmod with i
replace eff  with n
```

on remplit les champs de la base temporaire asgqltmp.dbf qu'on trie par ordre décroissant via `index on -eff to tmp`. Le reste de la boucle tant que de calcul `do while fn > " "` vient stocker les effectifs des autres variables dans la base asgql.dbf et il ne reste plus qu'à trier et à afficher de façon propre, ce qui se fait dans la boucle tant que suivante (`do while nbm <= nmax`) à coups de fonction `str` et de concaténations de chaine comme dans

```
store msg+" "+m+" "+n+" ( "+p+" %)" to msg
```

On viendra donc là encore lire soigneusement le programme suivant.

```
*************************************************************
*                                                           *
*    ASGQL.prg : Analyse Statistique Générale               *
*                de variables QuaLitatives                  *
*                                                           *
*************************************************************

*** saisie du nom de la base

  CLEAR
  SET TALK OFF
  SET SAFETY OFF

  * dbase oblige à initialiser les variables de get, donc :

  store dbf() + replicate(" ",20) to nombase
  store substr(nombase,1,20) to nombase

  @ 8,03 say " Nom de la base ? : "
  @ 9,03 say " (ou Aide pour explications)"
  @ 8,50 get nombase
  read
```

```
*** explications

  IF substr(upper(nombase+"    "),1,4) = "AIDE"
  .OR. len(trim(ltrim(nombase)))=0
     clear
     ? "Allql.prg : ASG, toutes variables QL "
     ?
     ? " Vous devez disposer d'une base de données au sens Dbase"
     ? " (fichier de type .DBF) dont la structure doit être la suivante :"
     ?
     ? "   - un premier champ de type Caractère de dimension 4"
     ? "   - tous les autres champs sont numériques. "
     ?
     RETURN
  ENDIF

*** test de la présence de la base

  store trim(ltrim(nombase)) to nombase
  IF at(".DBF",upper(nombase)) = 0
    store nombase+".DBF" to NOMComplet
  ELSE
    store nombase to NOMComplet
    store at(".DBF",upper(nombase)) to ip
    store substr(nombase,1,ip-1) to nombase
  ENDIF

 IF .NOT. FILE(NOMComplet)
   ? chr(7)
   ? chr(7)
   ? " Désolé, je ne vois pas ce fichier. Vérifiez son existence avec DIR "
   ? " Ne tapez pas .DBF mais indiquer le chemin d'accés ( PATH ) "
   ? " Fin anormale d'exécution, code 1 : fichier non trouvé."
   ?
   RETURN
 ENDIF

*** partie véritable des calculs

set talk   off
set safety off
set alter  to asgql.sor
set alter  on
```

```
? " (gH) - asgql.prg ; Angers, 1995"
? "     Analyse statistique générale de variables qualitatives"
? "     pour la base ",DBF(),"   ;    ",DATE(),TIME()

store dbf() to nomorg
count to nbt
store 1 to c
store field(c) to fn
use asgqltmp
zap
use asgql
zap
do while fn > " "

   if .not. type(fn)="N"
      ? "   ",fn," variable numéro ",c,"  non numérique"
   else

      * détection des codes min et max

      index on &fn to tmp
      goto  top
      store &fn to imin
      goto  bottom
      store &fn to imax

      * passage en revue de tous les codes
      * stockage dans asgqltmp

      use asgqltmp
      zap
      store imin to i
      do while i <= imax
         use &nomorg
         count for &fn=i to n
         use asgqltmp
         append blank
         replace nmod with i
         replace eff  with n
         store i+1 to i
      enddo
```

```
      * tri des effectifs et stockage dans asgql

      use     asgqltmp
      index   on -eff to tmp
      count   to nbm
      goto    top
      store   recno() to il

      use     asgql
      append  blank
      replace nomvar with fn
      replace nbmod  with nbm
      store   recno() to jl

      use     asgqltmp index tmp
      goto    il
      store   0 to nu
      do      while .not. eof()
              store nmod to nmd
              store eff  to veff
              use   asgql
              goto  jl
              store nu+1 to nu
              store 2+2*nu to ndc
              store field(ndc)   to nc
              store "replace "+nc+" with "+str(nmd) to cmd
              &cmd
              store ndc+1        to ndc
              store field(ndc)   to nc
              store "replace "+nc+" with "+str(veff) to cmd
              &cmd
              use asgqltmp index tmp
              goto il
              skip
              store recno() to il
      enddo
   endif
   use &nomorg
   store c+1 to c
   store field(c) to fn
enddo
```

```
* tri selon le mode et affichage
use asgql
index on  -eff01 to tmp
? "  NOM  Mode  Effectif          Modalité  Effectif"
goto top
do while .not. eof()
   store nomvar to msg
   * il y a 20 modalités au plus, le nb exact est dans nbmod
   store nbmod to nmax
   store 1  to nbm
   do while nbm <= nmax
      store field(2+2*nbm) to  nnmod
      store &nnmod       to  nmod
      store str(nmod,6)  to m
      store field(3+2*nbm) to  neff
      store &neff        to  eff
      store str(eff,6)   to n
      store str(eff*100/nbt,5,2)  to p
      * on peut aller jusqu'à 7 variables par ligne ...
      * mais souvent on ne veut pas dépasser 3
      store 3 to vdr
      store msg+"   "+m+"   "+n+"    ( "+p+" %)"   to msg
      store nbm+1 to nbm
      if mod(nbm,vdr)=0
         ? "   ",msg
         store replicate(" ",len(nomvar)) to msg
      endif
   enddo
   if len(trim(msg)) > 0
      ? "   ",msg
   endif
   skip
enddo
use &nomorg
?
? " -- fin de asgql "
?
close alter
?
? " Résultats dans Asgql.sor"
?
set talk   on
set safety on
```

4.5 Le logiciel Excel

Excel est au départ un tableur c'est à dire un logiciel gestionnaire de cases (ou "cellules"). Une cellule peut contenir une valeur numérique ou caractère tapée directement dans la cellule[3], une formule précédée par le signe =, un tableau, un graphique, un bouton déclenchant une macro (programme *Excel*).

Pour la partie qui nous intéresse, à savoir l'analyse statistique, *Excel* présente de nombreux points intéressants. Tout d'abord les onglets qui permettent de donner dans un même fichier plusieurs rubriques de natures différentes comme le descriptif, les données, les analyses statistiques, les commentaires. Ensuite les formules, simples à écrire. Par exemple pour sommer les valeurs des cellules `B2, B3...B19`, il suffit de mettre dans une cellule l'expression `=SOMME(B2:B19)`[4]. *Excel* réalise même automatiquement ces instructions si on sélectionne la plage désirée et si on clique sur le bouton Σ.

De nombreuses fonctions sont disponibles, et pas seulement statistiques. Leur appel, avec les paramètres, les restrictions d'utilisation s'apprend assez facilement. A l'aide du menu Insérer sous-menu Fonction on fait apparaitre les catégories de fonctions ce qui permet de choisir la fonction. Celle-ci sélectionnée, un assistant de fonction propose de saisir les divers paramètres et compose automatiquement l'expression, les paramètres étant séparés par des points-virgules.

Excel 5 propose les catégories de fonctions suivantes

```
BASE DE DONNEES           DATE ET HEURE
FINANCES                  INFORMATION
LOGIQUE                   MATHS ET TRIGO
RECHERCHE ET MATRICES     STATISTIQUES
TEXTE
```

L'aide sur les fonctions statistiques est souvent minimale, suffisante pour un(e) statisticien(ne), inutilisable pour un(e) débutant(e) en statistiques et en *Excel*, le rappel des formules n'aidant pas forcément à comprendre ce qu'il faut en faire ni à quoi elles servent.

3. pour entrer 1 en tant que caractère, il faut taper `="1"` dans la cellule
4. les débutants en *Excel* auront intérêt à essayer aussi `=SOMME(B2;B19)` qui n'effectue que la somme des deux cellules `B2` et `B19`. Le symbole ; indique une liste ou concaténation de valeurs, le symbole : une plage de cellules.

Les fonctions statistiques disponibles en *Excel* 5 sont

```
AVERAGEA
BETA.INVERSE
CENTILE
CENTREE.REDUITE
COEFFICIENT.ASYMETRIE
COEFFICIENT.CORRELATION
COEFFICIENT.DETERMINATION
COVARIANCE
CRITERE.LOI.BINOMIALE
CROISSANCE
CROISSANCE
DROITEREG
ECART.MOYEN
ECARTYPE
ECARTYPEP
ERREUR.TYPE.XY
FISHER
FISHER.INVERSE
FREQUENCE
GRANDE.VALEUR
INTERVALLE.CONFIANCE
INVERSE.LOI.F
KHIDEUX.INVERSE
KURTOSIS
LNGAMMA
LOGREG
LOI.BETA
LOI.BINOMIALE
LOI.BINOMIALE.NEG
LOI.EXPONENTIELLE
LOI.F
LOI.GAMMA
LOI.GAMMA.INVERSE
LOI.HYPERGEOMETRIQUE
LOI.KHIDEUX
LOI.LOGNORMALE
LOI.LOGNORMALE.INVERSE
LOI.NORMALE
LOI.NORMALE.INVERSE
LOI.NORMALE.STANDARD
LOI.NORMALE.STANDARD INVERSE
LOI.POISSON
LOI.STUDENT
LOI.STUDENT.INVERSE
LOI.WEIBULL
MAX
MEDIANE
MIN
MODE
MOYENNE
MOYENNE.GEOMETRIQUE
MOYENNE.HARMONIQUE
MOYENNE.REDUITE
NB
NBVAL
ORDONNEE.ORIGINE
PEARSON
PENTE
PERMUTATION
PETITE.VALEUR
PREVISION
PROBABILITE
QUARTILE
RANG
RANG.POURCENTAGE
SOMME.CARRES.ECARTS
TENDANCE
TEST.F
TEST.KHIDEUX
TEST.STUDENT
TEST.Z
VAR
VAR.P
```

Pour une analyse univariée de variables quantitatives, des fonctions comme `MOYENNE`, `MIN`, `MAX`, `VAR`, `VAR.P`, `ECARTYPE` et `ECARTYPEP` fonctionnent sur le modèle de la fonction `SOMME` précédente à savoir : `=FONCTION(PLAGE)`. On ira lire l'aide sur `VAR`, `VAR.P`, pour retrouver la différence entre le calcul de la variance d'une population en se basant sur la population entière et l'estimation de cette variance en se basant sur un échantillon de la population. La poignée de recalcul d'*Excel* permet de faire glisser une formule de calcul d'une colonne à l'autre sans avoir ni à entrer la fonction ni la plage de calculs.

Au niveau de l'analyse bivariée de ces mêmes variables quantitatives, la fonction de base est `COEFFICIENT.CORRELATION`. Pour les deux séries de valeurs `B2...B19` et `C2...C19` on calcule leur coefficient de corrélation par `=COEFFICIENT.CORRELATION(B2:B19;C2:C19)`. Les valeurs de a et b pour le modèle linéaire $Y = aX + b$ s'obtiennent par `=DROITEREG(y;x)` où x et y sont soit une plage de valeurs soit leur son nom. Attention toutefois à bien sélectionner deux cellules et à exécuter la fonction en mode matriciel (en faisant *Shift-Control-Enter* au lieu de *Enter*) car sinon *Exce*l ne renvoie que la première valeur calculée, celle de a[5].

Les nombreux graphiques d'*Excel* permettent facilement de tracer les courbes, points et autres indicateurs visuels qui permettent d'apprécier tous ces calculs.

En ce qui concerne les variables qualitatives, la fonction de base est le comptage (tri à plat ou tri croisé) qui s'obtient à partir du menu Données, sous-menu Rapport de Tableau Croisé Dynamique. Le tri à plat est obtenu en ne précisant que la variable pour les lignes, en spécifiant `NBVAL` comme calcul, et le tri à plat s'obtient en donnant un nom de variable pour les lignes et un nom de variable pour les colonnes. Les options de calcul permettent d'afficher les comptages absolus et relatifs, les totaux par ligne, par colonne. *Excel* permet au passage d'autres calculs multicritères comme la moyenne pour un couple de modalités données... Là encore, les nombreux graphiques d'*Excel* permettent de visualiser rapidement ces résultats.

Au niveau des variables textuelles, *Excel* ne propose aucun outil en standard. Il faudrait découper les phrases en mots, à raison de un mot par cellule pour commencer à pouvoir traiter les variables textuelles.

5. au pire, on peut utiliser `=INDEX(DROITEREG(y;x;"VRAI");1)` pour obtenir a et `=INDEX(DROITEREG(y;x;"VRAI");2)` pour avoir b.

On se méfiera également d'*Excel* dans les affichages :*Excel* peut avoir des formats d'affichages surprenant pour des non-initiés[6]. Il faut signaler aussi que les résultats des formules ne sont pas considérées comme des valeurs, ce qui pose parfois des problèmes pour trier des moyennes, des variances et il faut alors recourir à une technique de Copier / Collage Spécial - Valeurs pour obtenir l'affichage désiré. De même, les résultats affichées dans les comptages sont toujours par défaut par ordre croissant des valeurs et il y a donc quelques manipulations à effectuer pour aboutir à nos *affichages intelligents*.

4.6 Programmation statistique sous Excel

Excel se programme avec *VBA* (Visual Basic pour Applications). Montrons par l'exemple comment effectuer des calculs pour des variables quantitatives. A partir d'un classeur vide, on passe en mode éditeur de macros avec la menu Outils sous-menu Macros option Macros (ou la combinaison de touches Alt-F8). On nomme AstQT la macro dans la ligne prévue à cet effet et on appuie sur le bouton Créer. On se retrouve alors dans l'environnement *Microsoft Visual Basic* avec comme titre de fenêtre [Module1(Code)] où le texte

```
Sub AsgQT()
End Sub
```

a déjà été écrit automatiquement. Une fois le reste du texte écrit, testé, documenté, on peut enregistrer le fichier *Excel* comme modèle, permettant ainsi un usage des macros pour d'autres fichiers. Admettons que notre macro ait été enregistrée dans le fichier modèle `AsgQt.xlt`. Une utilisation possible de ce fichier est de l'ouvrir, d'y incorporer les données à traiter, de l'enregistrer sous forme de classeur puis de lancer la macro avec cette fois-ci la commande Exécuter au lieu de la commande Créer. Nous avons prévu que notre macro travaille en trois temps : d'abord une détection du nombre de lignes et de colonnes présentes, puis un calcul univarié et enfin le calcul de la matrice des corrélations. Nous utiliserons donc 3 sous-programmes désignés par DimDossier, Univariee et MatCorr.

6. nous avons vu récemment dans une feuille les 5 cellules suivantes

1	1	1	somme	2

en fait, les 3 premières cellules contenaient 0,6 mais le format de cellule était d'afficher les valeurs en nombres entiers.

Excel laisse la possibilité à l'utilisateur d'écrire les données n'importe où. Notre hypothèse sera que les données sont mises dans un bloc rectangulaire qui commence en ligne 1, colonne 1. Ainsi pour les données

DONNEES	TRN	AVN	HTL	Cmp	Glf	Tns
Ain	5	2	4	1	2	6
Aisne	3	0	5	0	1	8
Allier	0	1	3	1	0	9
Alpes HP	1	4	5	2	0	5
Alpes H	4	2	3	4	0	12
Alpes M	2	4	0	6	1	1
Eure	8	1	0	5	1	2
Finistère	1	3	2	2	0	3
Gard	4	2	1	3	1	4
...						
Yonne	6	1	2	6	0	6

notre programme détectera comme données le bloc des lignes nommées AIN, AISNE...EURE pour les colonnes TRN, AVN, HTL. Notre critère de détection est donc que les données sont dans des cellules non vides. Commençant avec `NbLig = 1` la boucle

```
While Not (Cells(NbLig, 1) = "")
```

vient donc trouver le nombre de lignes de données. Pour le nombre de colonnes, nous avons utilisé la même idée mais avec l'instruction

```
While IsEmpty(Worksheets(1).Cells(1, Nbcol).Value) = False
```

Cette instruction, qui n'est valide que si les données sont dans la feuille (ou onglet) numéro 1, est mise uniquement pour montrer comment on accède aux onglets. On aurait aussi pu mettre `Worksheets("Feuille1")` si l'onglet devait être nommé ainsi, recopier les données dans l'onglet "Calculs Stats" et utiliser alors `Worksheets("Calculs Stats")` ou tout autre endroit au gré du programmeur...

Nous avons décidé de mettre nos résultats à la suite des données, utilisant la variable `FinLig` pour savoir à quelle ligne écrire. Une vrai application viendrait certainement créer des onglets pour l'analyse univariée des variables quantitatives, pour l'analyse univariée des variables qualitatives, etc. mais notre macro n'est qu'un exemple.

Une formule comme `=MOYENNE("B2:B6")` qu'on entrait avec le clavier ou à la souris doit maintenant être définie comme `Application.Average(r)` où r est la bonne plage de données à traiter (le passage du terme français au terme anglais est un choix imposé par $Microsoft$). Pour la variable i, cette plage est `Range(Cells(2, i), Cells(Finlig, i))` et il est raisonnable d'afficher la valeur correspondante arrondie à 10^{-2} près, à la ligne numéro colonne 2 par l'instruction

```
Worksheets(1).Cells(Finlig + 2, i) = Application.Round(m, 2)
```

De la même façon, pour la matrice des corrélations, deux boucles imbriquées avec l'appel de la fonction `Application.Correl(r1, r2)` suffisent.

Voici donc le texte complet de la macro, qu'on lira avec soin, avec d'abord le programme principal puis les sous-programmes.

```
Sub AsgQT()
' ******************************************
' *
' * Exemple de macro (programme) en Excel :
' *    calcul de moyenne, écart-type etc.
' *
' ******************************************
'  on appelle en fait les trois sous-programmes
'     DimDossier, Univariee et MatCorr
' # Détermination du nombre de lignes et de colonnes
    DimDossier NbLig, Nbcol, Finlig
' # calcul de moyennes, écart-type et coefficients de variation
    Univariee NbLig, Nbcol, Finlig
' # matrice des corrélations
    MatCorr NbLig, Nbcol, Finlig
End Sub
```

```
' *******************************************
'*
         Sub DimDossier(NbLig, Nbcol, Finlig)
'*
' *******************************************

' # 1.1 Détermination du nombre de lignes

NbLig = 1
While Not (Cells(NbLig, 1) = "")
    NbLig = NbLig + 1
Wend
Finlig = NbLig   ;    NbLig = NbLig - 1
Cells(Finlig + 1, 1) = " Dossier avec : "
Cells(Finlig + 1, 2) = NbLig
Cells(Finlig + 1, 3) = " lignes"
Cells(Finlig + 1, 4) = " et"

' # 1.2 Détermination du nombre de colonnes

Let Nbcol = 1
While IsEmpty(Worksheets("Feuil1").Cells(1, Nbcol).Value) = False
    Let Nbcol = Nbcol + 1
Wend
Let Nbcol = Nbcol - 1
Worksheets(1).Cells(Finlig + 1, 5) = Nbcol
Worksheets("Feuil1").Cells(Finlig + 1, 6).Formula = " colonnes "
Finlig = Finlig + 1
End Sub

' *******************************************
'*
         Sub Univariee(NbLig, Nbcol, Finlig)
'*
' *******************************************

' # 2. calcul de moyennes, écart-type et coefficients de variation

Cells(Finlig + 2, 1) = "Moyenne"
Cells(Finlig + 3, 1) = "Ecart-type"
Cells(Finlig + 4, 1) = "cdv"
Cells(Finlig + 5, 1) = "min"
Cells(Finlig + 6, 1) = "max"
```

```
For i = 2 To Nbcol
    r = Range(Cells(2, i), Cells(Finlig, i))
    m = Application.Average(r)
    s = Application.StDevP(r)
    p = Application.Min(r)
    g = Application.Max(r)
    Worksheets(1).Cells(Finlig + 2, i) = Application.Round(m, 2)
    Worksheets(1).Cells(Finlig + 3, i) = Application.Round(s, 2)
    Worksheets(1).Cells(Finlig + 4, i) = "?"
    Worksheets(1).Cells(Finlig + 5, i) = Application.Round(p, 2)
    Worksheets(1).Cells(Finlig + 6, i) = Application.Round(g, 2)
    cdv = -1
    If Abs(m) > 0 Then
        cdv = 100 * s / m
    End If
    Worksheets("Feuil1").Cells(Finlig + 4, i) = Application.Round(cdv, 1)
Next i

End Sub

' *******************************************
'*
         Sub MatCorr(NbLig, Nbcol, Finlig)
'*
' *******************************************

' # 3.matrice des corrélations

Finlig = Finlig + 8
Worksheets(1).Cells(Finlig, 1) = " Matrice des corrélations "

' Labels des lignes et colonnes de la matrice

For i = 2 To Nbcol
    Worksheets(1).Cells(Finlig + 1, i) = Worksheets(1).Cells(1, i)
Next i
For i = 2 To Nbcol
    Worksheets(1).Cells(Finlig + i, 1) = Worksheets(1).Cells(1, i)
Next i
```

```
' Calculs des coefficients de corrélation

For i = 2 To Nbcol
    r1 = Range(Cells(2, i), Cells(Finlig, i))
    For j = 2 To i
      r2 = Range(Cells(2, j), Cells(Finlig, j))
      rho = Application.Correl(r1, r2)
      Worksheets(1).Cells(Finlig + i, j) = Application.Round(rho, 2)
    Next j
Next i

End Sub
```

4.7 Awk et les analyses textuelles

Si *Dbase* et *Excel* peuvent convenir pour des analyses de données numériques, il n'en va pas de même avec les variables textuelles, ne serait-ce qu'à cause du découpage des phrases en mots. Certains logiciels comme *Alceste*, *Spad.T* (Système Portable d'Analyse des Données Textuelles) sont disponibles sur le marché, mais nous allons présenter ici une solution moins couteuse (même si elle ne fait que des analyses élémentaires) avec le langage *Awk*.

Le langage de progammation *Awk*, nommé d'après l'initiale de ses auteurs Alfred V. Aho, Brian W. Kernighan et Peter J. Weinberger est un petit langage de manipulations de fichiers texte, soits d'après ses auteurs *"Awk is a convenient and expressive programming language that can be applied to a wide variety of computing and data-manipulation tasks"*. La plupart des implémentations de Awk sont des interpréteurs qui lisent le programme, traitent les données dans la foulée et affichent les résultats (ou les mettent dans des fichiers). Ce langage est disponible gratuitement sous licence GNU et est alors nommé gawk. Il est disponible sur PC, UNIX etc. Il a sa propre *FAQ*, son groupe de *news* etc.

Awk a été écrit dans le but de réaliser du parcours de fichiers textes, apellé aussi "filtrage actionnel" et il est donc trés performant pour cela.

Quand on traite des fichiers, un programme a souvent la structure classique suivante :

```
DEBUT d'algorithme
  ouvrir le fichier Don
  tant que non fin de fichier sur Don
       lire la ligne ChdeCarc sur Don
       si CONDITION sur ChdeCar
                 alors TRAITEMENT de ChdeCar
       finsi CONDITION sur ChdeCar
  fintant que non fin de fichier sur Don
  fermer Don
FIN d'algorithme
```

Le langage *Awk*, a été conçu pour faciliter ce genre de programme et traite le même fichier avec la seule structure :

```
CONDITION { TRAITEMENT }
```

Ici, le fichier *Don* est sous-entendu (et se déduit de l'appel de *Awk*). Chaque ligne *ChdeCar* de *Don* est passée en revue de façon automatique. Si besoin est, le mot-clé BEGIN signifie "à exécuter avant d'ouvrir le fichier", le mot END "après parcours du fichier". Chaque ligne du fichier quand elle est lue devient la ligne courante et est notée \$0. Elle possède NF champs nommés \$1,\$2... \$NF (variables de noms réservés mais de contenu modifiable). Ces champs sont les "mots" d'une "phrase" avec des séparateurs "normaux" (espace blanc, virgule, etc.).

AWK peut être utilisé en mode direct par

AWK ' instructions ' données

ou en mode programme par la séquence d'appel :

AWK -f programme données

Il faut noter que données peut être ambigu (par exemple DOC*.*) ; il y a alors exécution du fichier programme sur chacun des fichiers correspondants. Le filtrage élémentaire ne retient que les lignes comportant un certain texte, mise entre *slashs*.

Par exemple, si on veut avec *Awk* afficher toutes les lignes (avec leur numéro de ligne dans le fichier) qui contiennent le mot "Auteur" dans tous les fichier de type .Dic du répertoire courant, on peut écrire directement en session Dos ou Unix ou OS/2 etc. :

```
AWK '/Auteur/ { print FILENAME,FNR,\$0 }' *.Dic
```

Pour un travail plus conséquent, on ou mettre l'instruction /Auteur/ print FILENAME,FNR, $0 et la suite dans le fichier Aut.Awk et on viendra alors taper

```
AWK -fAut.Awk *.Dic
```

Awk tire sa concision d'une syntaxe élémentaire proche du langage *C*, de ses variables prédéfinies, comme

- $0 : la ligne courante toute entière
- NF : son nombre de champs (ou de mots)
- $i : son champ (ou mot) numéro *i* (si *i* est numérique)
- NR : le numéro de la ligne courante dans le fichier
- FILENAME : le nom du fichier en cours

et de son mécanisme de gestion des expressions régulières.

Rappelons qu'une expression régulière est une chaîne de caractères éventuellement répétés, générée à partie des règles suivantes :

- tout caractère est expression régulière (notée E.R. en abrégé)
- la concaténation de deux E.R. est une E.R.
- la répétition d'une E.R. (ou la chaîne vide) est une E.R. ;
  répétée 0 fois ou plus, elle est notée *, répétée 1 fois ou plus, notée +
  répétée 0 fois ou une (ou la chaîne vide), elle est notée?
- l'alternance de deux E.R. qui est notée | et qui signifie "ou" est une E.R.

Awk permet de tester le début de chaîne par ^, la fin de chaîne par $, n'importe quel caractère par . ; une classe de symboles se note entre crochets [ ]. Certains caractères sont réservés, à savoir ^ $. [ ] | ( ) * + ? ; leur équivalent caractère est accessible via un *antislash* supplémentaire Dans une classe, ^ signifie la négation de l'expression régulière exprimée.

Des expressions régulières simples (et leur signification) sont par exemple

`^B`	chaine qui commence par B
`^G$`	chaîne qui finit par G
`^.$`	chaîne à un seul caractère
`^[AEIOU]`	chaîne avec une des voyelles de la classe
`^[ABC]`	chaîne avec A,B ou C en début
`^[^a-z]$`	chaîne à un seul caractère qui n'est pas une lettre minuscule

Mais *Awk* gère des expressions plus techniques, telles `^[0-9]+$` qui détecte une chaîne non nulle avec seulement des chiffres, `^(\+|-)?[0-9]+\.?[0-9]*` qui correspond à un nombre réel avec signe éventuel, point éventuel, décimales éventuelles.

La condition de test sur une ligne en cours de programme peut se résumer à l'instruction

```
/expression_régulière/  { opérations à effectuer }
```

ou être plus complexe. Le symbole de validité de modèle (ou "pattern matching") est . En fait, /expr/ est un résumé pour /expr/ $0. Par défaut, l'instruction associée est print $0, c'est-à-dire affichage de toute la ligne. Par exemple, pour afficher toutes les lignes d'un fichier contenant les lettres A,I et R dans cet ordre, avec éventuellement des caractères entre ces lettres, il suffit d'écrire `/*A.*I.*R/ { print FILENAME,NR,\$0 }`.

Awk dispose aussi d'un mécanisme d'indexation dynamique, ce qui lui permet de gérer des "tableaux en mémoire associative". Cela signifie simplement que les indices dans les tableaux sont libres, non limités aux entiers. Ainsi, tab[mot]++ (où mot est une chaîne de caractères quelconques) permet de comptabiliser les mots d'un texte dans le tableau noté tab.

Awk est un langage non typé, initialisant. Toute variable utilisée prend une valeur par contexte. Les valeurs considérées comme numériques sont initialisées à zéro. Astucieusement, comme nous l'avons déjà dit, la syntaxe de base de *Awk* est celle du langage *C*. Cela s'explique par la volonté de ses auteurs de promouvoir *Awk* comme un langage d'un abord facile et préparant au C[7] ou permettant de l'éviter dans les cas simples.

7. le deusième auteur du langage Awk cité précédemment est bien sûr le père du C avec D. M. Ritchie.

Après cette longue introduction, regardons le tout petit programme *Awk* suivant où les lignes qui commencent par # sont des commentaires :

```
# Dico.Awk

BEGIN { fs = "tmp.tmp" }

(FNR==1) { print " Traitement des mots dans : " FILENAME }

{  nbl ++                              # comptage des lignes
   gsub(/["',.%#/]/ , " ")             # filtrage des séparateurs
   gsub(/[();:-]/, " ")                # élmination de la ponctuation
   for (i=1;i<=NF;i++) { ++mot[ $i ] } # comptage des mots
   nbm += NF }

END {  for (x in mot) { printf("%-40s %6d\n", x , mot[x]) > fs }
       close(fs)     }
```

Appliqué à un texte, il met dans le fichier tmp.tmp le dictionnaire des mots, comme par exemple :

```
offrant                                      1
22_03_19XX                                   1
ai                                           5
apprentissage                                4
déclinent                                    1
sont                                        15
bien                                        16
oeuvré                                       1
informe                                      3
les                                         51
longuement                                   1
expliquant                                   1
pu                                           1
```

qu'on peut trier[8] par ordre alphabétique, par ordre décroissant d'occurrence.

8. compte-tenu de la "particularité" de certains langages (le programme *Awk* est lui-même écrit en *C* et donc en "américain" et ne gère pas les accents), nous laissons le tri à un module externe, une commande "maison"... plutot que de faire ce tri en *Awk*.

On obtient alors respectivement le fichier alphabétique

```
21_h                                      1
22_03_19XX                                1
23_11_19XX                                1
4e_technologique                          1
abandonné                                 1
abord                                     1
abruti                                    1
absence                                   3
académie                                  1
accepte                                   2
accepté                                   2
```

et le fichier décroissant d'occurrences.

```
de                                      212
la                                      122
le                                      119
et                                      113
l                                        90
à                                        83
d                                        77
que                                      65
un                                       64
du                                       63
il                                       52
ce                                       19
mère                                     19
```

En ce qui concerne les condordances, la recherche d'environnement, il n'est pas très difficile d'écrire un petit programme *Awk* qui cherche si le mot i correspond à celui demandé et qui vient afficher les mots $i-2$, $i-1$, i, $i+1$ etc. la seule difficulté venant de l'affichage à réaliser...

4.8 Le logiciel Addad

En Analyse des données, on doit souvent relancer les mêmes programmes avec des paramètres différents, que ce soit pour afficher des résultats complémentaires, pour travailler avec une partie seulement des données, pour utiliser des identificateurs avec plus de caractères etc.

La chaine logicielle *ADDAD* utilise, à part des fichiers de données (fichiers textes) des fichiers de paramètres qui viennent adapter l'exécution des programmes aux besoins des utilisateurs. Pour constituer un fichier de travail ADDAD, il faut écrire des ordres stricts avec une syntaxe rigoureuse dans un fichier ascii lisible. On peut l'écrire directement sous éditeur ou, (mieux) modifier un ancien fichier de paramètres. Chaque ordre commence par le symbole $. Après ce symbole, on trouve une instruction de controle (RUN, END) ou un paramètre (FORMAT, PRT, ...). Il faut vérifier soigneusement que $ figure en colonne 1 et que chaque ordre est correctement écrit. De plus, les commandes sont à écrire dans l'ordre suivant :

```
$RUN  $PON  $Lxxx   $F05
TITRE PARAM OPTIONS GRAPHE LISTE
$PRT  $F11  $F...   $END
```

Ainsi, pour exécuter une AFC (Analyse Factorielle des Correspondances) on indique d'abord qu'il faut exécuter le module ANCORR (ANalyse des CORRespondances) avec l'ordre $RUN ANCORR (le fichier exécutable sur le disque est ANCORR.EXE). La commande $PON indique ensuite que les paramètres seront affichés à l'écran (sinon, mettre $POF).

Ensuite, la commande $L80 demande une sortie sur 80 caractères (si on sait utiliser des tout petits caractères, mettre $L132). Le reste des indications commence par la commande $F05=. car nous mettrons tous les paramètres dans le même fichier.

On indique ensuite le TITRE symbolique de l'analyse ; c'est une phrase qui sera reproduite sur chaque page du listing de sortie. Ainsi la ligne

```
TITRE  Exemple pour une AFC  par l'ADDAD ;
```

fera écrire Exemple pour une AFC par l'ADDAD sur chaque page de sortie. Le mot TITRE est un paramètre ou "mot-clé", le ; sert de terminateur au programme d'éxécution.

La description des autres paramètres de fait ensuite par la ligne de mot clé PARAM. On y indique le nombre d'individus (NI), le nombre de variables (NJ), le nombre de facteurs désirés (NF) avec un maximum de 7 facteurs affichés. On peut ici préciser le nombnre d'éléments à mettre en supplémentaire. Exemple de syntaxe :

```
PARAM NI=9 NJ=6 NF=5 LECIJ=1;
```

Viennent ensuite les OPTIONS de sortie des données, valeurs et vecteurs propres, des résultats et des graphiques. Cela commence par le mot clé : OPTIONS. Une valeur 1 signifie à écrire. La variable IOUT indique si on sort le tableau de données en entrée, IMPFI et IMPFJ sert pour les résultats Individus (I) et colonnes (J). NGR commande le nombre de graphiques. Si NGR est supérieur à 0 on décrit chaque graphique par l'instruction de mot clé GRAPHE. Par exemple :

```
OPTIONS IOUT=1 IMPVP=0 IMPFI=1 IMPFJ=1 NGR=2 ;
GRAPHE X=1 Y=2 GI=3 GJ=3 OPT=3 CADRE=1       ;
GRAPHE X=1 Y=3 GI=3 GJ=3 OPT=3 CADRE=1       ;
```

On ne détaille pas ici GI, GJ et les options de cadrage des graphiques (OPT,CADRE). Voir le manuel ADDADpour cela.

Enfin vient la description des colonnes, qui commence par le mot-clé LISTE. On met les noms des colonnes, séparés par des espaces blancs, et en donnant le cadrage entre parenthèses suivant le format (colonne de début, longueur de colonne).

On termine la liste des colonnes par un point-virgule. Exemple :

```
LISTE IDEN(1,4) CHROMO(10,1) CHINOIS(12,1) KLEE(14,1)
VAN_GOGH(16,1) DUFY(18,1) ROUSSEAU(20,1) ;
```

L'identificateur (mot clé IDEN obligatoire) commence à la colonne 1 et s'étend sur 4 caractères, la variable 1 s'appelle CHROMO et commence à la colonne 1à, elle s'étend sur 1 caractère etc. Le fichier de sortie des résultats est indiqué comme l'endroit où imprimer par $PRT=. On y met en général un nom de fichier *DOS* ou CON pour l'écran (peu conseillé) ou PRN pour l'imprimante. Exemple : $PRT=demoafc.sor Enfin, dans la plupart des cas, on mettra le nom fichier des données, précédé de l'instruction $F11=. (à cause de LECIJ=1, option par défaut). Exemple : $F11=DEMO.OK

D'autres spécifications de fichiers sont :

```
$PRT=demoafc.sor    (fichier de r\'{e}sultats)
$F21=demoafc.fal    (facteurs sur I)
$F22=demoafc.fal    (facteurs des \'{e}lts. sup sur I)
$F23=demoafc.fac    (facteurs sur J)
$F24=demoafc.fac    (facteurs des \'{e}lts. sup sur J)
```

La dernière ligne de ce fichier de travail contient impérativement : $END. Voici donc un exemple complet de fichier ADDAD pour réaliser une *AFC*.

```
$RUN    ANCORR
$L132
$PAR=.
TITRE    Exemple pour une AFC par l'ADDAD;
PARAM    NI=9 NJ=6 NF=4 LECIJ=1 STFI=1 STFJ=1;
OPTIONS  IOUT=2 IMPFI=1 IMPFJ=1 NGR=2;
GRAPHE   X=1 Y=2 GI=3 GJ=3 OPT=3 CADRE=1 ;
GRAPHE   X=1 Y=3 GI=3 GJ=3 OPT=3 CADRE=1 ;
LISTE    IDEN(1,4) CHROMO(10,1) CHINOIS(12,1) KLEE(14,1)
VAN_GOGH(16,1) DUFY(18,1) ROUSSEAU(20,1) ;
$PRT=  demoafc.sor
$F11=  demoafc.dat
$F21=  demoafc.fal
$F23=  demoafc.fac
$END
```

4.9 Le système Sas

En 1976, le mot *SAS* était l'acronyme de *statistical analysis system*. Depuis, *SAS* est une marque, celle du *SAS Institute*. Le logiciel *SAS* du début, logiciel d'analyse statistique est devenu un système universel de traitement de données, de base de données, de stockage de données... En France, on citera EDF et l'INSEE comme grands utilisateurs de *SAS* ; aux Etats-Unis, *SAS* est omniprésent dans l'administration, par exemple pour tout ce qui concerne la validation médicamentale thérapeutique (pour la *Foods and Drugs*...).

SAS fournit aujourd'hui un logiciel vraiment multi-plateforme : les exemples d'utilisation des fichiers lié au système d'exploitation déclinent systématiquement la syntaxe pour *MVS*, *VM/SP*, *Dos*, *Windows*, *MacOs*, *Linux*. Les services fournis par *SAS* autres que statistiques sont mutiples : gestion de fichiers, développement d'interfaces, gestion de l'information multi-réseaux, aide à la prise de décision, systèmes d'informations géographiques...

SAS s'impose, comme *usine à gaz* pour tout ce qui touche l'analyse des grandes bases de données, les statistiques de tout niveau, la gestion des grands comptes et des administrations. Logiciel "lourd" (la version 6 prend plus de 500 Mégas d'espace disque), le système *SAS* peut tout, fait tout, au prix d'un apprentissage et d'une location couteuse. Car *SAS* ne s'achète pas, il se loue (en fait, on paie pour obtenir un mot de passe, sinon *SAS* refuse d'effecteur les calculs [9].

SAS s'utilise en mode interactif via 3 fenêtres (*Program editor/F5*, *Log/F6* et *Output/F7*. ou en mode programme. Les instructions *SAS* sont fort limitées (il y a proc, data et options principalement). La richesse de *SAS* est dans le nombre de procédures et dans leurs options. Par exemple la procédure CLUSTER qui effectue des classifications hériarchiques propose 11 méthodes différentes avec 6 paramètres et 23 options...

Pour utiliser *SAS* en interactif, il faut écrire les instructions dans la fenêtre *Program editor*, les soumettre par *F3*, lire le détail des actions effectuées ou ratées dans la fenêtre *Log*, visualiser le cas échéant les résultats dans la fenêtre *Output* [10]. Il faut mettre RUN pour éxéuter les instructions. Chaque instruction, écrite éventuellement sur plusieurs lignes, se terme par un point-virgule.

9. seuls des étudiants penseraient à changer la date du système sur leur PC pour utiliser *SAS* sans contrainte, se privant ainsi de la gestion des fichiers liée au temps.

10. Une fois soumis, le texte dans Editor disparait mais on peut le rappeler par la touche *F4* (ou la commande *Rappeler le texte* du menu *Local*).

Par exemple, les deux lignes

```
proc options ;
run ;
```

génère un texte d'environ 350 lignes, qui donne le détail de toutes les options utilisées par de SAS avec leur valeur (options 150 lignes d'options dites portables, 200 lignes d'options dites de *HOST*) après l'en-tête, typiquement

```
NOTE:
  Copyright (c) 1989-1996 par SAS Institute Inc., Cary, NC, USA.
NOTE: SAS System SAS (r) Version 6.12
  Sous contrat de licence pour XXXX, Site XXXXXXXXXX.
```

La forme courte, à savoir `proc options`; provoque, quant à elle, la sortie de 150 lignes dont les suivantes:

```
OPTIONS PORTABLES :

NOBATCH BUFNO=3 BUFSIZE=0 BYERR BYLINE NOCAPS NOCARDIMAGE
CATCACHE=0 CBUFNO=0 CENTER NOCHARCODE CLEANUP NOCMDMAC
COMPRESS=NO CPUID DATE NODBCS DBCSLANG= DBCSTYPE=PCIBM NODETAILS
DEVICE= DFLANG=ENGLISH DKRICOND=ERROR DKROCOND=WARN NODMR DMS
NODMSBATCH DSNFERR NOECHOAUTO ENGINE=V612 NOERRORABEND
ERRORCHECK=NORMAL ERRORS=20 FIRSTOBS=1 FMTERR FMTSEARCH=
FORMDLIM= FORMS=DEFAULT GWINDOW HELPENV=WIN
HELPLOC=w:\sas\french\winhelp NOIMPLMAC INITCMD= INITSTMT=
INVALIDDATA=. LABEL LINESIZE=93 MACRO MAPS='!sasext0\maps'
MAUTOSOURCE MERROR MISSING=. NOMLOGIC NOMPRINT NOMRECALL
NOMSGCASE MSGLEVEL=N NOMSTORED MSYMTABMAX=4194304 NOMULTENVAPPL
MVARSIZE=4096 NEWS= NOTES NUMBER OBS=MAX NOOVP PAGENO=1
PAGESIZE=58 PARM= PARMCARDS=FT15F001 NOPRINTINIT PROBSIG=0 PROC
REMOTE= REPLACE REUSE=NO NORSASUSER S=0 S2=0
SAMPLOC=w:\sas\core\sassamp
```

```
OPTIONS HOST:

ALTLOG= ALTPRINT= AUTHSERVER= AUTOEXEC= AWSCONTROL=(SYSTEMMENU
MINMAX TITLE) AWSDEF=( 0 0 92 100) AWSMENU AWSMENUMERGE AWSTITLE=
CODEGEN COMAMID=TCP COMAUX1= COMAUX2= COMDEF=(BOTTOM CENTER)
CONFIG=W:\sas\CONFIG.SAS ECHO= EMAILDLG=NATIVE EMAILID= EMAILPW=
EMAILSERVER= EMAILSYS=MAPI NOEXITWINDOWS FONT=(Sasfont 10)
FONTALIAS= NOFULLSTIMER HONORAPPEARANCE HOSTPRINT NOICON
NOLOADLIST LOG= NUMKEYS=12 NUMMOUSEKEYS=3 PATH=w:\sas\CORE\WINNT
PATHDLL=w:\sas\french\sasdll PFKEY=WIN PRINT= PROCLEAVE=30K
REGISTER= RTRACE= RTRACELOC= SASCONTROL=(SYSTEMMENU MINMAX)
SET=(FT15F001 'FT15F001.DAT') SET=(SASEXT0 w:\sas) SET=(SASROOT
w:\sas) SET=(SASFOLDER "w:\sas") SET=(CBT101
!sasext0\sascbt\cbt101) SET=(CBT102 !sasext0\sascbt\cbt102)
SET=(CBT103 !sasext0\sascbt\cbt103) SET=(CBT104 SETCOMM=
SORTPGM=SAS SPLASH SPLASHLOCATION= STIMER SYSIN= SYSLEAVE=15K
SYSPRINT=(HP_Alex \\ALEX\HP LaserJet 5N) SYSPRINTFONT=
TOOLDEF=(TOP RIGHT) USERICON= NOVERBOSE WINCHARSET=ANSI XSYNC
XWAIT
```

Un programme SAS traditionnel vient définir les données avec l'étape DATA puis décrit les traitements à effectuer avec l'étape PROC. SAS sait bien sur lire dans tous les sens, tous les formats, que les fichiers soient formatés ou non, même s'il y a des données manquantes, que les individus soient sur une ou plusieurs lignes, qu'il y ait plusieurs individus sur une même ligne ou non, etc.

Comme premier exemple de programme, nous viendrons lire des données correspondant au dossier ELF, que nous allons trier par ordre de politesse[11]. L'instruction filename donne un nom symbolique au fichier des données, l'étape DATA utilise le paramètre infile pour lire le fichier, input précise le format d'entrée. La procédure sort effectue le tri selon le critère descending sexe descending age qui vient donc trier par ordre décroissant de sexe puis, en cas de *ex-aequo* par age décroissant et enfin on affiche les données avec print.

11. appelé aussi *ordre de guerre*, qui met les femmes et les plus jeunes d'abord, protégeant ainsi qui vous pensez...

Le programme complet qui utilise ces instructions est :

```
* tut07.sas ;

* définition des données ;

filename elf 'l:\gh\crs\stat\trysas\elf.dat' ;
data elf ;
 infile elf ;
 input ID $ 1-4 SEXE 9 AGE 12-13 PROF 17-18
       ETUD 23 REGI 28 ACTIO 33 ;

run ;

/* tri selon l'ordre de politesse ;
   femme d'abord, homme ensuite
   les vieux d'abord, les jeunes ensuite
   code sexe : 0=homme, 1=femme           */

proc sort data=elf  out=elf_Poli;
  by descending sexe descending age ;

* affichage sélectif dans l'ordre ;

proc print data=elf_Poli ;
  var id sexe age ;

* exécution ;

run ;
```

Le calcul de *moyenne*, *écart-type*, *coefficient de variation* se fait en passant des paramètres à la procédure means. Ainsi

```
proc means data=elf n mean cv maxdec=1; var age ;
```

affectue ces calculs pour la variable *age* avec une seule décimale à l'affichage.

SAS sait présenter les tris à plat par ordre décroissant de fréquence (proc freq order=freq ;), gérer des formats d'entrée (*informat*) et des formats de sortie pour donner de beaux affichages. Ainsi pour le tri croisé du couple (SEXE,ETUDE), le programme *SAS*

```
proc format ;
 value sexf  0='Homme' 1='Femme' ;
 value etudf 0='NR' 1='Primaire' 2='Bepc' 3='Bac' 4='Sup' ;

filename elf 'l:\gh\crs\stat\trysas\elf.dat' ;
data   elf_beau ;
infile elf ;
input  ID $ 1-4 SX 9 AGE 12-13 PROF 17-18
       ETUD 23 REGI 28 ACTIO 33 ;
label  SX='Sexe de la personne' ;
label  ETUD='Niveau d''études' ;
format SX sexf. ETUD etudf. ;

proc freq data=elf_beau ;
 table etud*sx / norow nocol nopercent ;

run ;
```

produit les résultats suivants

```
          TABLE DE ETUD PAR SX

ETUD(Niveau d'études)     SX(Sexe de la personne)
```

Fréquence	Homme	Femme	Total
NR	2	1	3
Primaire	1	5	6
Bepc	7	23	30
Bac	8	13	21
Sup	17	22	39
Total	35	64	99

4.10 Programmation statistique en Sas

Passons maintenant à un exemple plus conséquent. Reprenant le dossier $Vins$, nous décidons d'afficher les résultats statistiques par moyenne décroissante puis par ordre de coefficient de variation décroissant. Pour cela, nous stockons avec OUT les résultats de la procédure MEANS sans rien afficher (NOPRINT). Le fichier de résultat ne se présentant pas sous une forme convenable pour notre tri, nous le transposons avec la procédure TRANSPOSE. Ensuite, nous renommons les colonnes avec un identifiant plus explicite (moy plutot que $col4$, etc.). et nous calculons à la main le cdv. Il ne reste plus qu'à trier par ordre décroissant (SORT... BY DESCENDING...) et à afficher pour obtenir ce que l'on veut ! Voici donc le programme complet :

```
libname cours 't:\' ;
data asg_tmp ; set cours.vins ;
proc means data=asg_tmp noprint ; output out=tmp_gsa ;
proc transpose data=tmp_gsa output=tmp_asg ;
data vinsasg ;
  set tmp_asg ;
  if col1=0 then delete ;
  if _name_='_FREQ_' then delete ;
  min = col2 ;
  max = col3 ;
  moy = col4 ;
  ect = col5 ;
  if moy > 0 then cdv = 100*abs(ect)/abs(moy) ;
             else cdv = -1 ;
proc sort  data=vinsasg ;  by descending moy ;
proc print data=vinsasg ;  var _name_ moy ect cdv min max ;
proc sort  data=vinsasg ;  by descending cdv ;
proc print data=vinsasg ;  var _name_ moy ect cdv min max ;
run ;
```

Dans le même genre d'idées, SAS calcule la matrice des corrélations et l'affiche toujours en mode rectangulaire. Puisqu'il s'agit d'une matrice symétrique, l'usage est d'en effectuer un affichage intelligent via la matrice triangulaire supéreure correspondante. SAS ne fournit en standard aucune procédure qui effectue ce travail. Montrons en quelques lignes comment réaliser cet affichage.

On commence d'abord par définir la variable LESQT qui contient le nom de toutes les colonnes numériques (nous supposons ici que toutes les variables sont quantitatives, sinon, on laisse le lecteur mettre la liste des seules variables quantitatives) et on calcule la matrice des corrélations (proc corr) sans affichage(noprint) sur cette liste de variables (var &lesqt;). La procédure rollout vient alors définir *outp* comme "ensemble de données" (*dataset* dans le jargon de *SAS*) ne retenant que la partie "CORR" des résultats ; elle utilise la liste lesqt comme tableau de caractères et une boucle pour col de 1 à nblignes réalise l'affichage.

```
/* by courtesy of Bill Raynor */
%let lesqt='_NUMERIC_' ;
proc corr data=vins noprint outp=outp ;
   var &lesqt ;
run ;

data rollout ;
   set outp ;
   where (_TYPE_ eq 'CORR') ;
   array c {*} &lesqt ;
   l = dim(c) ;
   row = _N_ ;
   do col = 1 to row ;
      corr = c{col} ;
      output ;
     end ;
   keep row col corr ;
 run ;
```

Nous renvoyons le lecteur à un cours de *SAS* plus approfondi pour maitriser les mutliples options de *SAS* au niveau de l'étape DATA, pour découvrir les procédures de calcul accessibles via l'étape PROC, le langage des macros de *SAS*, la notion de référence, de librairie...

Donnons pourtant un dernier exemple qui illustre la force de *SAS*. Nous avions précédemment montré comme faire une analyse univariée de variables quantitatives, avec un tri par moyenne décroissante puis par coefficient de variation décroissant. Hormis la partie DATA ces instructions sont suffisament générales pour s'appliquer à n'importe quelles données. Si ces instructions sont mises dans le fichier asgqt.sas, en modifiant légèrement le début pour que les donnéees soient asg_tmp, soit la ligne :

```
proc means data=asg_tmp noprint ; output out=tmp_gsa ;
```

alors n'importe quelle base de variables quantitatives (nommons la DONNEES) peut être traitée grâce aux instructions suivantes

```
data asg_tmp ; set DONNEES ;
 %include 'asgqt.sas' ;
```

On conçoit bien que *SAS*, avec ses programmes textes interprétés par le système soit portable au niveau du code. Pour les données, c'est beaucoup plus technique. *SAS* utilise *en local* un système de stockage lié au système d'exploitation mais offre des moyens de convertir les données en un "format de transport", moins compact mais portable.

De plus *SAS* permet de développer des applications complètes avec menus déroulants, interfaces utilisateurs etc. L'utlisation de *SAS* peut donc se faire à tous les niveaux : simple utilisateur de commandes et touches à pousser, lignes de commandes "à la volée", écriture de programmes complexes...

Arrivé(e) à un certain niveau de statistiques, que ce soit à cause des tailles de données ou de la complexité des opérations statistiques à effectuer, *SAS* devient incontournable, malheureusement au gout de certain(e)s...

4.11 Les autres logiciels statistiques

Il n'est pas possible de présenter tous les logiciels statistiques disponibles. Toutefois, il est possible de donner quelques repères quant au choix du logiciel à utiliser (et éventuellement à acheter). Le premier point est la qualité des manuels fournis avec le logiciel. En particulier, les logiciels qui fournissent des exemples d'utilisation détaillés (comme *StatView*) et qui rappellent les formules utilisées (comme *Sas*) sont beaucoup plus agréables à utiliser que les manuels de référence seule.

Le second point à prendre en compte est la facilité des fonctions d'import et d'export. Le standard .dbf de *Dbase*, en particulier, doit être présent pour garantir une portabilité maximale des données et de leur structure. Le troisième point est la qualité des graphiques disponibles, leur intégration au document final de synthèse de l'analyse statistique.

Enfin, le dernier point est l'aide à l'interprétation à la rédaction, les garde-fous de calcul et de résultats... A ce titre, deux logiciels se dégagent de tous les autres : ce sont SAS et Statistica. En effet, SAS et Statistica fournissent tous deux des aides aux calculs, des référénces écrites sur les calculs, éventuellement leurs auteurs, les conditions d'application.

Par exemple dans la version 6 de SAS, la procédure *CALIS* vient effectuer une *Covariance Analysis of LInear Structural equations* c'est à dire une analyse de la covariance des équations linéaires structurales (!). Un chapitre entier du volume 1 du manuel SAS/STAT User's Guide : c'est le chapitre 14. Il comporte 120 pages au format A4 et contient une cinquantaine de références bibliographiques sur le sujet, la plus ancienne étant celle de *Wilson* en 1931 et la plus récente celle de *McDonald* en 1989. On est loin d'une demi-page d'aide sur la syntaxe des paramètres pour une commande sous *Excel*...

Chapitre 5.

Conclusion

Il y a au moins deux points de vue fondamentaux sur les statistiques : on peut les considérer comme une science pure, "LA" statistique avec ses modèles, ses équations, ses variables aléatoires et ses modèles, et on peut la considérer comme un ensemble de techniques pour appréhender le monde, avec ses calculs, ses imprécisions inhérentes au domaine d'étude, ses interprétations sur les résultats...

Il serait vain de vouloir restreindre les statistiques à l'un ou l'autre des points de vue car ils sont complémentaires. Comment appliquer en effet une formule si on ne sait pas à quoi elle correspond, et même, si elle s'applique ? A quoi sert, aussi, de connaitre une formule si on ne s'en sert pas, si on ne sait pas l'utiliser ? [1]

Théorie et pratique sont donc nécessairement liées. Et malheureusement la pratique ne peut se faire aujourd'hui sans les ordinateurs. D'où un biais, celui du moule qu'impose le logiciel choisi. C'est pourquoi il nous a semblé important de présenter les mêmes calculs avec divers logiciels, d'insister sur les méthodes. Car dans 10 ans de nombreux logiciels auront changé, les formules seront toujours les mêmes et les utilisateurs seront toujours en train de naviguer entre les formules et les logiciels...

1. la réponse : savoir une formule sert à passer l'examen n'a aucun sens pour le monde réel, celui des entreprises, des applications.

BIBLIOGRAPHIE

A. Blanchet, A. Gotman
L'enquête et des méthodes : l'entretien
Natan Université, Paris 1992.

J.L. Boursin
Comprendre les statistiques descriptives
Armand Colin, 1988.

G. Celeux, E. Diday, G. Govaert
Classification automatique des données
Dunod informatique, 1989.

R. P. Cody, J. K. Smith
Applied statistics and the sas programming language
Prentice-Hall, 1997.

L. D. Delwiche, S. J. Slaughter
The little sas book, a primer
Sas Institute Inc., 1995.

F. de Singly
L'enquête et des méthodes : le questionnaire
Natan Université, Paris 1992.

Y. Dodge
Statistique, dictionnaire encyclopédique
Dunod, Paris 1993.

Y. Evrard, B. Pras, E. Roux
Market : études et recherches en marketing, fondements et méthodes
Nathan, 1993.

L. Lebart, A. Morineau, J.P. Fénelon
Traitement des données statistiques
Dunod, 1982.

L. Lebart, A. Morineau, M. Piron
Statistique exploratoire multidimensionnelle
2^{eme} édition, *Dunod, 1997.*

L. Lebart, A. Salem
Analyse Statistique des données textuelles
Dunod, 1988.

B. Scherrer
Biostatistique
Gaetan Morin éditeur, 1984.

M. Tenenhaus
Méthodes statistiques en gestion
Dunod entreprise, 1994.