Gilles HUNAULT

1999

Probabilités et Statistiques

Université d'Angers

Table des matières

Chapitre 1.

Introduction

1.1 Présentation générale

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 pour donner une vue globale, tentent de synthétiser ou de modéliser en se servant des modèles théoriques issus des probabilités.

On trouvera ici à la fois une introduction aux calculs mathématiques qui servent de base aux probabilités et aux statistiques mais aussi des exemples concrets, une discussion de l'interaction entre calculs et interprétation. Nous fournirons aussi des exemples de programmation de façon à rendre le lecteur, la lectrice à la fois autonome quant à la lecture d'ouvrages sur les deux disciplines qui fondent ce cours mais aussi capable de s'adapter aux logiciels courants ou apte à (re)programmer ces calculs et formules.

Nous attendons des élèves de ce cours qu'ils sachent calculer (car on ne "fait" pas des mathématiques) à la fois à la main et avec un ordinateur. On espère qu'ils sauront maitriser l'ensemble des concepts et qu'ils pourront acquérir au long de ces pages suffisamment de recul pour "voir des choses" au-delà des pages de formules et de chiffres, par-delà les hypothèses, les modèles. Enfin, et surtout, nous voudrions qu'ils arrivent à exposer leurs vues, leur résultats chiffrés, leurs conclusions avec précision, concision et culture en un tout lisible, accessible dans cet effort difficile qu'est la rédaction et l'exposé.

1.2 Plan de l'ouvrage

Les pages que vous êtes en train de lire constituent le chapitre 1, c'est l'*Introduction* à l'ouvrage. Le chapitre 2 viendra rappeler les bases du calcul des *Probabilités*, avec de nombreux exemples concrets. On y introduira les *Variables Aléatoires*, leurs grandeurs caractéristiques et quelques *Lois* classiques. On passera en revue dans le chapitre 3 les *Statistiques* classiques et leurs caractéristiques avec un petit détour par la notion de test. Ensuite, des statistiques un peu moins classiques seront traitées dans le chapitre 4 avec des *variables autres* que qualitatives ou quantitatives. Enfin, la théorie reprendra sa place avec les *variables vectorielles* (en particulier *gaussiennes*) et la *convergence* dans le chapitre 5 alors que le chapitre 6 essaiera de conclure via l'*informatisation* des traitements, l'utilisation de *logiciels* et l'importance de la *rédaction*.

Sachant que tous et toutes n'ont pas les mêmes bases théoriques et mathématiques, nous avons souvent regroupé les démonstrations en fin de chapitre de façon à faciliter la lecture des concepts et à ne pas ralentir la progression du cours. Attention toutefois aux démonstrations : de façon à privilégier la réflexion, les démonstrations sont incomplètes. On ne trouve aucune justification des enchainements. C'est tout à fait volontaire, contraignant et sans doute plus pédagogique qu'il n'y parait. Le lecteur peu intéressé par les mathématiques nous en saura gré, l'expert se fera un plaisir de justifier ici l'utilisation de la commutativité de l'addition dans $\mathbb{R}$, là l'existence des intégrales généralisées... Dans le même esprit, des annexes sur les structures algébriques et le dénombrement complètent les 6 chapitres et avec aussi quelques incontournables tables numériques.

De nombreux exercices sont proposés après chaque partie de cours importante. Certains sont des applications pas toujours simples du cours, d'autres viennent fournir des compléments, des éclairages différents sur les notions mises en place. Nous ne saurions trop conseiller de lire soigneusement les énoncés de ces exercices, à défaut de les résoudre...

Chapitre 2.

Espaces Probabilisés et Variables aléatoires

2.1 Probabilités

2.1.1 Définitions et Propriétés élémentaires

Soit E un ensemble dont les éléments sont nommés *évènements élémentaires* ; E lui-même est nommé *espace des épreuves*. On le note classiquement Ω. On appelle *évènement* de E tout élément de $\mathcal{T} = \mathcal{P}(E)$. Le complémentaire de l'évènement F par rapport à E est noté $\overline{F}$ et nommé évènement *contraire* de F.[1] Une probabilité p sur E est une fonction de $\mathcal{T}$ dans $[0,1]$ qui vérifie les axiomes suivants nommés *axiomes de Kolmogorov* :

[P1] $p(E) = 1$
[P2] $p(A \cup B) = p(A) + p(B)$ si $A \cap B = \emptyset$

Tout évènement X tel que $p(X) = 1$ est appelé *évènement certain*. Tout évènement Y tel que $p(Y) = 0$ est appelé *évènement impossible*. En particulier, E est certain et $\emptyset$ est impossible. $(E,\mathcal{T})$ est nommé espace probabilisable et $(E,\mathcal{T},p)$ est nommé espace probabilisé.

1. On note aussi $\mathcal{C}(F)$ ou comp(F) ce complémentaire suivant le contexte (mathématique, informatique...). De même, on précise aussi par rapport à quoi on prend le complémentaire s'il peut y avoir ambiguité, comme avec la notation traditionnelle $\complement_E F$.

De **[P1]** et **[P2]** on déduit que

[P3] $p(\emptyset) = 0$

[P4] $p(\overline{F}) = 1 - p(F)$

[P5] $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ (formule de *Poincaré*)

[P6] p est croissante de $(\mathcal{T}, \subset)$ dans ([0,1], $\leq$)

$\mathcal{T}$ est nommé l'univers des possibles (ou plus simplement *univers*). Au lieu d'évènement, on emploie aussi le terme d'éventualité. Deux événements A et B tels que $A \cap B = \emptyset$ sont dits *incompatibles* (ou encore disjoints). Deux événements A et B tels que $p(A \cap B) = p(A).p(B)$ sont dits *indépendants*. Ces deux notions ne sont pas liées. Un *système complet* d'évènements de E est une partition de E.

2.1.2 Probabilités conditionnelles, décompositions

On appelle probabilité conditionnelle de B quand A la quantité

$$p(B|A) = \frac{p(A \cap B)}{p(A)}.$$

Cette notion n'a bien sûr de sens que si $p(A)$ est non nul. On prendra garde au fait que $p(B|A)$ n'est *qu'une notation* et qu'elle n'a aucun sens "classique" puisque $B|A$ ne désigne pas un évènement au même titre que $A \cap B$, $A \cup B$...[2]

Grâce aux propriétés ci-dessus, pour deux évènements A et B, on a :

la *formule des probabilités totales*

$$p(A) = p(B).p(A|B) + p(\overline{B}).p(A|\overline{B})$$

et la *formule de Bayes*

$$p(B|A) = \frac{p(B).p(A|B)}{p(B).p(A|B) + p(\overline{B}).p(A|\overline{B})}$$

2. $A \subset B$ non plus n'est pas un ensemble. Mais il n'y a ici aucune confusion puisqu'on n'écrit jamais $p(A \subset B)$.

Ces formules se généralisent, pour un système complet $\{\omega_i\}$ de E en

[P7] $p(A) = \Sigma_i \; p(\omega_i).p(A|\omega_i)$
[P8] $p(\omega_j|A) = p(\omega_j).p(A|\omega_j)/\Sigma_i \; p(\omega_i).p(A|\omega_i)$

Pour A fixé, notons maintenant $p_{|A}$ l'application $X \mapsto p(X|A)$. Alors :

[P9] $p_{|A}$ est une probabilité.

2.1.3 Généralisation de la notion de probabilité

Une probabilité est en fait un cas particulier de mesure sur une tribu. Une *tribu* (ou σ-algèbre ou algèbre σ-additive) sur un ensemble Ω est une famille de sous-ensembles de Ω qui contient $\emptyset$ et Ω, stable par passage au complémentaire et par union finie ou dénombrable. Si $\mathcal{T}$ est une tribu sur Ω, le couple $(\Omega,\mathcal{T})$ est appelé *espace mesurable.* Une *mesure* μ sur un espace mesurable est une fonction *σ-additive* de Ω dans $\overline{R}_+$ (où $\overline{R}_+$ désigne $R_+ \cup \{\infty\}$) c'est à dire une fonction qui vérifie : $\mu(\bigcup_i \; A_i) = \Sigma_i \; A_i$ où les A_i sont une famille finie ou dénombrable d'évènements deux à deux incompatibles. Le triplet $(\Omega,\mathcal{T},\mu)$ est alors appelé *espace mesuré.*

2.1.4 Calculs effectifs de Probabilités

Les définitions et propriétés précédentes montrent comment calculer la probabilité d'un évènement à partir d'une décomposition ensembliste en évènements élémentaires dont on connait les probabilités. Celles-ci sont soit déterminées par des calculs de dénombrements soit en utilisant une hypothèse raisonnable, pratique parce que simple mais pas forcément fondée : la répartition équitable ou équirépartition. Ainsi obtenir un nombre pair en lançant un dé se ramène à l'union disjointe de l'obtention d'un deux, d'un quatre... Sauf hypothèse contraire, tous les chiffres sont équiprobables et sortir un chiffre pair a donc une probabilité de 0.5.

Les probabilités ainsi construites sont toutes *théoriques* en ce sens qu'elles ne reflètent pas la *réalité*. Un dé n'a jamais ses faces parfaitement équiprobables, le "**hasard**" intervient plus souvent qu'un "ordre" uniformément réparti...

2.1.5 Quelques calculs de Probabilités

Si n tomes d'une encyclopédie sont rangés sur une étagère, la probabilité que les tomes 1 et 2 soient rangés cote à cote dans cet ordre est $(n-1).(n-2)!$ soit $(n-1)!$. De même, la probabilité que les tomes 1,2...t soient rangés cote à cote dans cet ordre est $(n-t+1)!$.

Dans un polygone convexe déterminé par n points, il y a $n(n-1)/2$ droites dont n cotés. La probabilité qu'une droite soit une diagonale dans un tel polygone est donc $(n-3)/(n-1)$.

Prenons $2r$ chaussures parmi n paires. La probabilité de n'avoir pris aucune paire correcte est $2^{2r}C_n^{2r}/C_{2n}^{2r}$.

2 personnes écrivent au hasard un nombre de deux chiffres exactement (entre 10 et 99). Cette expérience est répétée n fois. La probabilité d'avoir écrit au moins une fois le même nombre est $1-(89/90)^n$.

r personnes sont dans un ascenseur à n étages. On suppose que la probabilité qu'une personne s'arrête à un étage donné est $1/n$ et que les décisions des r personnes sont indépendantes. La probabilité que les r personnes s'arrêtent à des étages différents est A_n^r/n^r et la probabilité que deux personnes exactement s'arrêtent au même étage et que les $r-2$ autres personnes s'arrêtent chacune à un étage différent (qu'on supposera différent de l'étage de ces deux personnes) est $C_r^2 A_n^{r-1}/n^r$.

Admettons qu'un "dé" A comporte 4 faces rouges et 2 faces blanches et qu'un "dé" B comporte 2 faces rouges et 4 faces blanches. On lance une pièce de monnaie et si on obtient pile, on utilise le dé A ; si c'est face, on utilise le dé B. La probabilité d'avoir utilisé A sachant qu'on a obtenu n fois rouge en lançant n fois le dé est $2^n/(2^n+1)$.

2.2 Variables aléatoires

2.2.1 Définitions

Maitenant que nous savons reconnaitre, décomposer et plus généralement manipuler des évènements, nous allons introduire des fonctions sur ces évènements. Ainsi, nous savons déterminer les probabilités associés aux différentes faces d'un dé, que ce soit dans des tirages avec ou sans remise, mais que faire avec ces tirages ? Par exemple, pour deux dés, on peut s'intéresser à la somme des chiffres obtenus en un tirage, à leur produit, ou au nombre de lancers qu'il faut effectuer pour obtenir le 6 sur chaque dé en même temps...

C'est ici qu'interviennent les variables aléatoires. Formellement, une variable aléatoire nommée X sur un ensemble E muni d'une probabilité p est une fonction de $\mathcal{T}_E$ dans une partie de $F = \mathbb{R}$ telle que $\forall u \in F, X^{-1}(B) \in \mathcal{T}_E$ où B est un *borélien*[3] de F. Si $Im(X)$ est fini ou dénombrable, le vocabulaire classique[4] l'appelle variable *discrète* ; sinon, elle est dite *continue*. On peut aussi prendre comme ensemble d'arrivée non pas $\mathbb{R}$ mais $\mathbb{R}^n$; la variable aléatoire est alors dite *vectorielle*. Un premier problème posé est alors celui des notations et un second celui de la probabilité sous-jacente à une "expérience aléatoire empirique". Pour le premier problème, on conviendra de noter $X \in B$ l'ensemble des ω tel que $X(\omega) \in B$ et on s'autorisera même à écrire $X = k$ si B se réduit à une seule valeur, alors que l'écriture correcte serait $X^{-1}(\{k\})$. On trouvera aussi un sens à une expression comme $X \leq k$, $a < X < b$... Pour le second problème, ce sera souvent plus délicat car il y de nombreux sous-entendus : si on lance deux dés et qu'on s'intéresse à la somme des valeurs obtenues, la probabilité sous-jacente n'est pas très difficile à trouver, même si elle requiert un peu de calcul.

Il est facile à partir de la notion de variable aléatoire de comprendre une différence fondamentale entre probabilités et statistiques. En probabilités, on fera l'hypothèse que chaque face du dé est *équiprobable*, on s'intéressera à toutes la valeurs possibles et on déterminera théoriquement ce qu'on doit obtenir.

3. Nous utilisons ce terme sans l'avoir défini, ce qui ne gêne pas si l'on prend comme équivalent la périphrase "réunion d'intervalles". L'intérêt de donner le "bon" mot est de faciliter l'approfondissement vers la théorie de la mesure...

4. On peut être étonné de ce vocabulaire puisque'une variable aléatoire est... une fonction ! De même pour la qualification variable aléatoire continue puisque ce n'est pas la fonction qui est continue mais l'ensemble d'arrivée qui est non dénombrable.

En statistiques, on réalise une expérience et on compte. C'est en comparant les valeurs observées et les valeurs théoriques qu'on peut se faire une idée exacte du phénomène, (se) convaincre que les hypothèses sont exactes, ou remettre en question le modèle (ici l'équiprobabilité).

Soit donc X une variable aléatoire discrète sur (E,p) à valeurs dans I et définissons l'application p_X de I dans $[0,1]$ par $p_X(k) = p(X^{-1}(\{k\}))$ soit encore, avec la notation précédente : $p_X(k) = p(X = k)$. Alors :

[V1] p_X est une probabilité appelée probabilité image de p par X.

Passons maintenant à une variable aléatoire continue Y à valeurs dans $\mathbb{R}$, et notons p_Y l'application qui à un intervalle $[a,b]$ de $\mathbb{R}$ associe la quantité $p(a \leq Y \leq b) = p(Y^{-1}([a,b]))$. Alors

[V2] p_Y est une probabilité.

Là encore, p_Y est appelée probabilité image de p par Y.

2.2.2 Valeurs caractéristiques

Essayons maintenant de caractériser nos variables aléatoires. Comme elle peuvent avoir de nombreuses valeurs, il est intéressant de les résumer. Le premier indicateur numérique global est la *moyenne* qui donne la *tendance centrale*. Si X est une variable aléatoire discrète, notons x_i les différentes valeurs possibles de X et p_i les probabilités associées. Alors la moyenne de X est $m = \Sigma\ x_i.p_i$.

Pour une variable aléatoire continue Y, comme il n'est pas possible de d'utiliser $p(x)$,[5], on a recours à la densité de Y qui est une fonction f_Y telle que $\displaystyle\int_{\mathbb{R}} f = 1$ avec $p(Y \leq y) = \displaystyle\int_{-\infty}^{y} f$. La moyenne de Y est alors $m = \displaystyle\int Id.f_y$.

Il existe diverses notations pour désigner la moyenne. Lorsqu'on utilise plusieurs variables aléatoires, on peut noter m_X, $m(X)$, $moy(x)$, $average(X)$, $av(X)$ ou encore $\overline{X}$ pour préciser de quelle variable il s'agit.

Le deuxième indicateur numérique global est la variance. C'est un indicateur de *dispersion absolue* moyenne autour de m. Si on note $mt(X,n) = \Sigma\ x_i^n.p_i$ le moment d'ordre n pour une variable aléatoire discrète X, alors on peut remarquer que $m(X) = mt(X,1)$.

5. Pourquoi ?

La variance peut alors être définie par $V = mt(X,2) - m(X)^2$. En d'autre termes, "la variance de X est la différence entre la moyenne du carré des valeurs et la carré de la moyenne des valeurs". On peut aussi introduire les moments centrés : $mt_c(X,n) = mt(X - m(X),n)$. On démontre alors que

[V3] $V(X) = mt_c(X,2)$

soit encore : "la variance de X est son moment centré d'ordre 2". Pour une raison peut-être pas très claire pour l'instant, on définit aussi l'écart-type $\sigma(X)$ de X comme la racine carrée de la variance : $\sigma(X) = \sqrt{V(X)}$. Comme pour la moyenne, de nombreuses notations existent pour $V(X)$ et $\sigma(X)$, notamment $var(X)$, $\overline{\overline{X}}$, $etyp(X)$, $ect(X)$, $std(X)$... Le rapport σ/m est nommé *coefficient de variation* et fournit un indicateur de *dispersion relative*.

Pour une variable aléatoire continue Y, le moment d'ordre n est défini par :

$$mt(Y,n) = \int_R Id^n.f_y$$

et la variance par

$$V = mt(Y,2) - m(Y)^2$$

Il existe d'autres indicateurs globaux (*kurtosis, skewness*) mais ils ne nous intéressent pas pour l'instant. Nous en parlerons au prochain chapitre quand nous traiterons les séries statistiques. De même, des indicateurs locaux (comme minimum, maximum...) ne sont pas a priori intéressants en probabilitiés. On verra par contre qu'ils prennent tout leur sens en statistiques et en informatique, à la fois comme indicateur et comme vérificateur de données.

Passons maintenant aux propriétés des opérateurs m, V et donc σ. m est additive c'est à dire que

[V4] $m(A + B) = m(A) + m(B)$

sous réserve que A et B soient deux variables aléatoires définies pour la même probabilité. m est aussi homogène soit :

[V5] $m(k.A) = k.m(A)$

où k est une constante. Ces deux propriétés font de m une *application linéaire*. Par contre la moyenne n'est pas multiplicative car $m(A.B) \neq m(A)m(B)$ en général (un contre-exemple suffit à le montrer). La variance n'est en général pas additive

[V6] $V(A + B) \neq V(A) + V(B)$

et non homogène puisque

[V7] $V(k.A) = k^2.V(A)$

On en déduit que σ n'est pas additive et que de plus

[V8] $\sigma(k.A) = |k|.\sigma(A)$

Le fait que V ne soit pas additive résulte du fait que m n'est pas multiplicative. En effet :

[V9] $V(A+B) - V(A) - V(B) = 2.(m(A.B) - m(A).m(B))$

On appelle covariance de A et B la quantité $m(A.B) - m(A).m(B))$ notée $cov(A,B)$ qu'on "normalise" pour obtenir le coefficient de corrélation linéaire entre A et B noté souvent ρ plutôt que $corr$:

$$\rho(A,B) = \frac{cov(A,B)}{\sigma(X).\sigma(Y)}$$

Les résultats suivants sont simples à démontrer

[R1] $Y = aX + b \Rightarrow |\rho(X,Y)| = 1$

[R2] $\forall A,B - 1 \leq \rho(A,B) \leq 1$

[R3] A et B indépendantes $\Rightarrow cov(A,B) = 0$

[R4] la relation $\mathcal{R}$ définie par
$x\mathcal{R}y \;\Leftrightarrow\;$ x et y sont en corrélation linéaire
est une relation d'équivalence.

Attention toutefois à ce que la réciproque que **[R3]** est fausse.

2.3 Lois Classiques

Les lois discrètes sont en général assez simples à aborder et le calcul de leurs caractéristiques (moyenne, écart-type) est assez facile à mener. Par contre, pour les lois continues et en particulier pour les dernières (Khi-deux, Student, Fisher-Snédécor) ces calculs demandent des techniques poussées d'intégration. Nous les citons cependant car elles sont importantes pour les test statistiques. De plus, elles font partie de la culture statistique et il est bon de les connaitre, au moins par leur nom et leur définition, si ce n'est par leurs formules.

2.3.1 Lois Discrètes

Loi de Bernoulli

C'est sans doute la loi discrète la plus simple à étudier puisqu'elle n'a que deux valeurs 0 et 1. Il ne faut pas pour autant la négliger puisqu'elle sert de support à tous les calculs sur les phénomènes binaires comme succés/échec, oui/non, présence/absence... On note traditionnellement $b(p)$ cette loi où p désigne la probabilité de $X = 1$. Sa moyenne est p, son écart-type $\sqrt{p.(1-p)}$.

Loi Binomiale

Cette loi requiert deux paramètres : n et p. On la note $\mathcal{B}(n,p)$. La façon la plus naturelle de présenter cette loi discrète est de la considérer comme la somme de n variables de Bernoulli indépendantes de même paramètre p. Sa moyenne est np et son écart-type $\sqrt{n.p.(1-p)}$. On peut aussi la présenter en termes de tirages de boules. Supposons qu'on tire avec remise n boules parmi un ensemble de boules de deux types (par exemple de deux couleurs, blanc et noir). Soit p la proportion de boules blanches. Alors $\mathcal{B}(n,p)$ peut s'interpréter comme la variable "nombre de boules blanches tirées". L'évènement "on a tiré k boules blanches" a donc comme probabilité $p_k = p(X = k) = C_n^k.p^k.(1-p)^{n-k}$ pour k de 0 à n.

Loi Uniforme Discrète

C'est là encore une loi très simple : elle est constante. On la définit classiquement sur l'intervalle $[\![1,n]\!] = [1,n] \cap \mathbb{N}$. Chaque probabilité élémentaire vaut donc $1/n$. La moyenne de $\mathcal{U}(n)$ est $(n+1)/2$ et son écart-type $\sqrt{(n^2-1)/12}$.

Loi Géométrique

Reprenons l'interprétation en termes de boules pour la loi Binomiale : on a n boules de deux couleurs, blanc et noir. Les boules blanches sont en proportion p. La loi Géomètrique $\mathcal{G}(n,p)$ correspond à la variable aléatoire X où $X = k$ est l'évènement "on a obtenu une (première) boule blanche seulement au k-ième tirage". La probabilité associée est $p_k = (1-p)^{k-1}.p$ d'où une moyenne de $1/p$ et un écart-type de $(1-p)/p^2$.

Loi de Pascal

Nommée aussi loi binomiale négative, cette loi généralise la loi géométrique. On s'intéresse à l'obtention de s boules blanches au bout de k tirages pour n boules de deux couleurs dont la proportion de boules blanches est p. La probabilité associée est $p_k = C_{k-1}^{s-1}.p^s.(1-p)^{k-s}$. La moyenne de $\mathcal{P}(s,p)$ est s/p et son écart-type de $\sqrt{s(1-p)/p^2}$.

Loi de Poisson

De paramètre λ, cette loi discrète est un peu surprenante puisqu'elle prend un nombre infini de valeurs à savoir 0,1,2.... Nommée aussi "loi des évènements rares", elle sert à modéliser les phénomènes de files d'attente. La probabilité d'obtenir le nombre entier k est $e^{-\lambda}\lambda^k/k!$. Sa moyenne est λ et son écart-type λ. On la note $\mathcal{P}(\lambda)$.

2.3.2 Lois Continues

Loi Uniforme Continue

Elle est constante, comme son nom l'indique. On la définit classiquement sur $[a,b]$. Sa moyenne est donc $(a+b)/2$ et son écart-type $\sqrt{(b-a)^2/12}$. On la note $\mathcal{U}([a,b])$.

Loi Triangulaire

Cette loi $\mathcal{T}([a,b])$ est définie sur $[a,b]$. Elle est composée de deux segments de droite qui se rencontrent au milieu de l'intervalle à une hauteur de $2/(b-a)$ ce qui dessine donc un triangle isocèle. On laisse en exercice la détermination de la probabilité élémentaire, le calcul de la moyenne qui vaut $(a+b)/2$ et de l'écart-type $(b-a)/2\sqrt{6}$.

Loi Exponentielle

Avec un paramètre α strictement positif, la loi $\mathcal{E}(\alpha)$ est définie sur $\mathbb{R}_+$ par la densité $f(x) = \alpha.e^{-\alpha x}$. Sa moyenne est $1/\alpha$ et son écart-type aussi.

Loi de Erlang-Gamma

Cette loi est définie par la densité $f(x) = \alpha^k.x^{k-1}.e^{-\alpha x}/\Gamma(k)$ pour $x \geq 0$, $\alpha > 0$ et $k > 0$ fixés, où Γ est bien sûr l'intégrale eulérienne $\Gamma(k) = \displaystyle\int_0^{+\infty} y^{k-1}.e^{-y}.dy$ qui généralise la fonction factorielle : $\Gamma(n) = (n-1)!$ pour n entier. La moyenne de la loi d'Erlang est k/α, son écart-type $\sqrt{k}/\alpha$.

Loi de Weibull-Rayleigh

La loi de Weibull est notée $\mathcal{W}(\alpha,\beta,\gamma)$. Elle est définie par la densité $f(x) = (\beta/\alpha).y^{\beta-1}e^{-y^\beta}$ pour $\gamma \leq x$, $\alpha > 0$, $\beta > 0$ et γ fixés, où $y = (x-\gamma)/\alpha$. On peut vérifier que α est un paramètre d'échelle, β un paramètre de forme et γ un paramètre de position. Sa moyenne est $\gamma + \alpha.\Gamma(1+1/\beta)$, son écart-type $\alpha.\sqrt{\Gamma(1+2/\beta) - \Gamma(1+1/\beta)}$. La loi de Rayleigh est $\mathcal{W}(\alpha,2,0)$.

Loi Normale

Cette loi à deux paramètres μ et ν a pour densité

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

où t est $(x-\mu)/\nu$. Sa moyenne vaut μ, son écart-type ν et on la note donc classiquement $\mathcal{N}(m,\sigma)$.

Loi de Cauchy

Si U et V sont deux lois normales centrées réduites, la variable $X = U/V$ est appelée loi de Cauchy. Elle n'admet ni moyenne ni écart-type.

Loi Log-Normale

Si U est une loi normale de moyenne μ et d'écart-type σ, la variable X définie par $X = e^U$ est appelée loi log-normale. Sa moyenne est $x^{\mu+\sigma^2/2}$, son écart-type $\sqrt{e^{2\mu+\sigma^2}e^{\sigma^2} - 1}$.

Loi du Khi-Deux

Si Z_1, Z_2 Z_ν... sont ν variables normales centrées réduites, alors la variable $\Sigma_i Z_i^2$ est une variable de Khi-deux à ν degrés de liberté. Sa moyenne est ν, son écart-type $\sqrt{2\nu}$.

Loi de Student

Si U est une variable normale centrée réduite, si V est une variable de Khi-deux à ν degrés de liberté et si U et V sont indépendantes, alors la variable $T = U/\sqrt{V/\nu}$ est une variable de Student à ν degrés de liberté. Sa moyenne est 0, son écart-type $\sqrt{\nu/(\nu-2)}$ pour $\nu > 2$.

Loi de Fisher-Snédécor

Si U est une variable de Khi-deux à ν_1 degrés de liberté et si V est une variable de Khi-deux à ν_2 degrés de liberté alors la variable $T = (U/\nu_1)/(V/\nu_2)$ est une variable de Fisher-Snédécor à (ν_1,ν_2) degrés de liberté. Sa moyenne est $\nu_2/(\nu_2 - 2)$ pour $\nu_2 > 2$.

2.4 Démonstrations des propriétés

• *Démonstration de la propriété* **[P3]**

D'après **[P2]** appliqué à $A = E$ et $B = \emptyset$ on a $p(E \cup \emptyset) = p(E) + p(\emptyset)$ car $E \cap \emptyset = \emptyset$. Puisque $E \cup \emptyset = E$, le membre de gauche est aussi égal à $p(E)$. On en déduit donc que $p(E) = p(E) + p(\emptyset)$. Dans R, puisque $p(E) = 1$, on obtient, en simplifiant par $p(E)$, le résultat recherché. Sous forme uniquement calculatoire, on peut réécrire cela en :

$$\begin{array}{llll} E \cup \emptyset = E & \Rightarrow & p(E \cup \emptyset) & = & p(E) \\ & \Rightarrow & p(E) + p(\emptyset) & = & p(E) \quad \text{car } E \cap \emptyset = \emptyset \\ & \Rightarrow & p(\emptyset) & = & 0 \ \diamondsuit \end{array}$$

• *Démonstration de la propriété* **[P4]**

D'après **[P2]** appliqué à $A = F$ et $B = \overline{F}$ on a $p(F \cup \overline{F}) = p(F) + p(\overline{F})$ car $F \cap \overline{F} = \emptyset$. Puisque $F \cup \overline{F} = E$, le membre de gauche est aussi égal à $p(E)$. On en déduit donc que $1 = p(F) + p(\overline{F})$ grâce à **[P1]**. Passant p(F) dans le membre de gauche on en déduit le résultat recherché. Sous forme d'équations :

$$\begin{array}{lllll} F \cup \overline{F} = E & \Rightarrow & p(F \cup \overline{F}) & = & p(E) \\ & \Rightarrow & p(F) + p(\overline{F}) & = & 1 \quad \text{car } F \cap \overline{F} = \emptyset \text{ et } p(E) = 1 \\ & \Rightarrow & p(\overline{F}) & = & 1 - p(F) \ \diamondsuit \end{array}$$

• *Démonstration de la propriété* **[P5]**

Nous aurons besoin du résultat suivant : $p(B \backslash C) = p(B) - p(C)$ si $C \subset B$ qu'on démontre via **[P2]** pour $B = C \cup (B \backslash C)$ puisque $C \cap (B \backslash C) = \emptyset$. Nous donnons la solution sous forme d'équations et d'implications, la justification des calculs étant laissée au lecteur. On a posé $D = B \backslash (A \cap B)$.

$$\begin{array}{llll} A \cup B = A \cup D & \Rightarrow & p(A \cup B) & = p(A \cup D) \\ & & & = p(A) + p(D) \\ & & & = p(A) + p(B \backslash (A \cap B)) \\ & & & = p(A) + p(B) - p(A \cap B) \ \diamondsuit \end{array}$$

• *Démonstration de la propriété* **[P6]**

Dire que p est croissante revient à dire que $A \subset B \Rightarrow p(A) \leq p(B)$. Or $A \subset B \Rightarrow B = A \cup (B \backslash A)$. En appliquant **[P1]**, on en déduit que $p(B) = p(A) + p(B \backslash A)$. Comme $\forall X p(X) \geq 0$, il vient $p(B) - p(A) \geq 0.\diamondsuit$

• *Démonstration de la propriété* **[P7]**

Commençons par démontrer la *formule des probabilités totales* qui en est un cas particulier. La définition de probabilité conditionnelle permet d'écrire "linéairement" $p(X \cap Y) = p(X).p(Y|X)$. Les calcus sont simples à mener :

$$\begin{array}{rcl} A & = & A \cap E \\ & = & A \cap (B \cup \overline{B}) \\ & = & (A \cap B) \cup (A \cap \overline{B}) \\ \text{d'où} \quad p(A) & = & p(A \cap B) + p(A \cap \overline{B}) \end{array}$$

et on conclut en remplaçant les deux termes du membre de droite par leur écriture "linéaire". Pour la formule générale, on applique la même méthode.

$$\begin{array}{rcl} A & = & A \cap E \\ & = & A \cap (\bigcup_i \ \omega_i) \\ & = & \bigcup_i \ (A \cap \omega_i) \\ \text{d'où} \quad p(A) & = & \Sigma_i \ p(A \cap \omega_i) \ \diamondsuit \end{array}$$

• *Démonstration de la propriété* **[P8]**

Commençons par démontrer la *formule de Bayes* qui en est un cas particulier. On commence par remarquer que $p(A \cap B) = p(B|A).p(A) = p(A|B).p(B)$. Puis :

$$\begin{array}{rcl} p(B|A) & = & p(A \cap B)/p(A) \text{ par définition} \\ & = & p(A|B).p(B)/p(A) \text{ compte-tenu de la remarque précédente} \\ & = & p(A|B).p(B)/p(A \cap B) + p(A \cap \overline{B}) \\ & & \text{grâce à la formule précédente des probabilités totales. } \diamondsuit \end{array}$$

Avec un système complet d'évènements, c'est à peine plus difficile :

$$\begin{array}{rcl} p(\omega_i|A) & = & p(A \cap \omega_i)/p(A) \\ & = & p(A|\omega_i).p(\omega_i)/p(A) \\ & = & p(A|\omega_i).p(\omega_i)/\Sigma_i \ p(A \cap \omega_i) \ \diamondsuit \end{array}$$

• *Démonstration de la propriété* **[P9]**

Il n'y a que les propriétés **[P1]** et **[P2]** à démontrer pour $p_{|A}$.

$$\begin{array}{rcl} p_{|A}(E) & = & p(E \cap A)/p(A) \\ & = & p(A)/p(A) \\ & = & 1 \ \diamondsuit \end{array}$$

donc la première propriété est démontréee.

De même :

$$\begin{array}{rcl} p_{|A}(X \cup Y) & = & p((X \cup Y) \cap A)/p(A) \\ & = & p((X \cap A) \cup (Y \cap A))/p(A) \\ & = & p(X \cap A)/p(A) + p(Y \cap A)/p(A) \\ & = & p_{|A}(X) + p_{|A}(Y) \ \diamondsuit \end{array}$$

et la deuxième propriété est aussi démontrée.

• *Démonstration de la propriété* **[V1]**

Remarquons tout d'abord que p_X n'est pas bien définie. Nous diposons des fonctions

$$\begin{array}{rcccl} p & : & \mathcal{P}(E) & \rightarrow & [0,1] \\ X & : & \mathcal{P}(E) & \rightarrow & I \\ p_X & : & I & \rightarrow & [0,1] \end{array}$$

Etendons la définition de p_X des singletons $\{k\}$ à un ensemble quelconque par $p_X(J) = \sum_{j \in J} p_X(\{j\})$ où $J \subset I$. Soient A et B sous-ensembles disjoints de I. Alors

$$\begin{array}{rcl} p_X(A \cup B) & = & \displaystyle\sum_{x \in A \cup B} p_X(\{x\}) \\ & = & \displaystyle\sum_{a \in A} p_X(\{a\}) + \sum_{b \in B} p_X(\{b\}) \\ & = & p_X(A) + p_X(B) \end{array}$$

De plus $p_X(I) = p(X^{-1}(I)) = p(E) = 1$ donc p_X est bien une probabilité.$\diamondsuit$

• *Démonstration de la propriété* **[V2]**

Là encore p_Y n'est définie que pour un intervalle. On l'étend naturellement pour tout *borélien*, c'est à dire pour toute réunion disjointe d'intervalles (notée $\dot\cup$ au lieu de $\cup$) par $p_Y(J) = \sum_{\dot\cup [a_i,b_i] = J} p_Y([a_i,b_i])$. Soient A et B sous-ensembles disjoints de $\mathbb{R}$. Alors

$$\begin{array}{rcl} p_Y(A \cup B) & = & \displaystyle\sum_{\dot\cup [x_i,y_i] = A \cup B} p_Y([x_i,y_i]) \\ & = & \displaystyle\sum_{\dot\cup [a_i,b_i] = A} p_Y([a_i,b_i]) + \sum_{\dot\cup [c_i,d_i] = B} p_Y([x_i,c_i]) \\ & = & p_Y(A) + p_Y(B) \ \diamondsuit \end{array}$$

• *Démonstration de la propriété* **[V3]**

$$\begin{aligned}
mt_c(X) &= moy(\ (X-m)^2\ ) \\
&= moy(X^2 - 2.m(X).X + m(x)^2) \\
&= moy(X^2) - 2.moy(X)m(X) + m(X)^2 \\
&= moy(X^2) - 2.m(X)^2 + m(X)^2 \\
&= moy(X^2) - m(X)^2 \\
&= V(X)\ \diamondsuit
\end{aligned}$$

• *Démonstration de la propriété* **[V4]**

Pour le cas discret :

$$\begin{aligned}
m(A+B) &= \Sigma(a_i + b_i).p_i \\
&= (\Sigma a_i.p_i) + (\Sigma b_i.p_i) \\
&= m(A) + m(B)\ \diamondsuit
\end{aligned}$$

Pour le cas continu :

$$\begin{aligned}
m(A+B) &= \int \Big(a(x) + b(x)\Big).p(x)dx \\
&= \int a(x).p(x)dx + \int b(x).p(x)dx \\
&= m(A) + m(B)\ \diamondsuit
\end{aligned}$$

• *Démonstration de la propriété* **[V5]**

Pour le cas discret :

$$\begin{aligned}
m(k.A) &= \Sigma k.a_i.p_i \\
&= k.\Sigma a_i.p_i \\
&= k.m(A)\ \diamondsuit
\end{aligned}$$

Pour le cas continu :

$$\begin{aligned}
m(k.A) &= \int k.a(x).p(x)dx \\
&= k.\int a(x).p(x)dx \\
&= k.m(A)\ \diamondsuit
\end{aligned}$$

La Démonstration de la propriété [**V6**] consiste à trouver un contre-exemple. Elle est laissée en exercice.

• *Démonstration de la propriété* [**V7**]

Pour le cas discret :

$$\begin{aligned} mt(k.A,2) &= \Sigma(k.a_i)^2.p_i \\ &= \Sigma k^2.a_i^2.p_i \\ &= k^2.\Sigma a_i^2.p_i \\ &= k.mt(A,2) \ \diamondsuit \end{aligned}$$

De même

$$\begin{aligned} m(k.A)^2 &= (k.m(A))^2 \\ &= k^2.m(A)^2 \end{aligned}$$

Donc finalement

$$\begin{aligned} V(k.A) &= mt(k.A,2) - m(k.A)^2 \\ &= k^2.mt(A,2) - k^2.m(A)^2 \\ &= k^2.(mt(A,2) - .m(A)^2) \\ &= k^2.V(A) \ \diamondsuit \end{aligned}$$

Pour le cas continu, même démarche puisque

$$\begin{aligned} mt(k.A,2) &= \int (\ k.a(x)\ )^2.p(x)dx \\ &= \int k^2.a(x)^2.p(x)dx \\ &= k^2.\int a(x)^2.p(x)dx \\ &= k.mt(A,2) \end{aligned}$$

• *Démonstration de la propriété* [**V8**]

$$\begin{aligned} \sigma(k.A) &= \sqrt{V(k.A)} \\ &= \sqrt{k^2.V(A)} \\ &= |k|.\sqrt{V(A)} \\ &= |k|.\sigma(A) \ \diamondsuit \end{aligned}$$

• *Démonstration de la propriété* [**V9**]

Comme :

$$\begin{aligned} mt((A+B),2) &= moy(\ (A+B)^2\ ) \\ &= moy(\ A^2+B^2+2.A.B\ ) \\ &= moy(A^2)+moy(B^2)+2.moy(A.B) \end{aligned}$$

et puisque :

$$\begin{aligned} moy((A+B)^2) &= (\ moy(A)+moy(B)\ )^2 \\ &= moy(A)^2+moy(B)^2+2.moy(A).moy(B) \end{aligned}$$

on en déduit

$$\begin{aligned} var(A+B)) &= mt(A+B,2)-m(A+B)^2 \\ &= moy(A^2)+moy(B^2)+2.moy(A.B) \\ &\quad \text{- } (\ moy(A)^2+moy(B)^2+2.moy(A).moy(B)\ ) \\ &= moy(A^2)-moy(A)^2+moy(B^2)-moy(B)^2 \\ &\quad +\ 2.(\ moy(A.B)-moy(A).moy(B)\ ) \\ &= var(A)+var(B)+2.cov(A,b)\ \diamondsuit \end{aligned}$$

• *Démonstration de la propriété* **[R1]**

Si $Y = aX + b$ alors

$$\begin{aligned} cov(X,Y) &= m(\ X.(a.X+b)\ )-m(X).m(a.X+b) \\ &= m(a.X^2+b.X)-m(X).(\ a.m(X)+b\ ) \\ &= a.m(X^2)+b.m(X)-a.m(X)^2-b.m(X) \\ &= a.(\ m(X^2)m(X)^2\ ) \\ &= a.V(X) \end{aligned}$$

Donc $\rho(X,Y) = cov(X,Y)/\sigma(X).\sigma(Y) = a.V(X)/|a|.\sigma(X).\sigma(X)$ et on a bien $|\rho(X,Y)| = |\ a/|a|\ | = 1.\diamondsuit$

La réciproque de la propriété peut s'énoncer ainsi :

$$|\rho(X,Y)| = 1 \Rightarrow \exists a,b\ ;\ m(\ (Y-(a.X+b))^2\ ) = 0.$$

Pour le montrer, considérons la variable $Z = m_X + \rho.(Y - m_Y).\sigma_X/\sigma_Y$ où ρ est une notation simplifiée de $\rho(X,Y)$. Notons aussi $\tau = \rho.\sigma_X/\sigma_Y$ pour simplifier les calculs. Alors

$$\begin{aligned} m(\ (Z-X)^2\ ) &= m(\ (X-m_X)^2\ )-2.\tau.m(\ (X-m_X).(Y-m_Y)\ ) \\ &\quad +\tau^2.m(\ (Y-m_Y)^2\ ) \\ &= V(X)-2.\tau.cov(X,Y)+\tau^2.V(Y) \\ &= V(X)-2.(\rho.\sigma_X/\sigma_Y).(\rho.\sigma_X.\sigma_Y)+(\rho.\sigma_X/\sigma_Y)^2.V(Y) \\ &= V(X)-2.\rho.\sigma_X^2+\rho.\sigma_X^2 \\ &= V(X).(1-\rho^2) \end{aligned}$$

Donc $|\rho| = 1 \ \Rightarrow\ m(\ (Z-X)^2\ ) = 0$. Cela signifie donc qu'en moyenne (quadratique) Z est ègal à X. Or $X = Z$ s'écrit $X = m_X + \rho.(Y - m_Y).\sigma_X/\sigma_Y$ soit encore $\sigma_Y.(X - m_X) = \rho.(Y - m_Y).\sigma_X.(Y - m_Y)$ ce qui donne la relation "linéaire" $Y = a.X + b$ avec $a = \rho.\sigma_Y/\sigma_X$ et $b = m_Y - a.m_x.\diamondsuit$

• *Démonstration de la propriété* **[R2]**

Soit Z la variable aléatoire $Z = X - m_X + \lambda.(Y - m_Y)$. Alors la quantité positive ou nulle $m(Z^2)$ est le trinome du second degré $T(\lambda) = a.\lambda^2 + b.\lambda + c$ avec :

$$\begin{array}{rclcl} a & = & m(\ (Y - m_Y)^2\ ) & = & V(Y) \\ b & = & 2.m(\ (X - m_X).(Y - m_Y)\ ) & = & 2.cov(X,Y) \\ c & = & m(\ (X - m_X)^2\ ) & = & V(X) \end{array}$$

Puique $T(\lambda)$ doit être de signe constant, son discrimant réduit est négatif, soit la relation :

$$cov(X,Y)^2 \leq V(X).V(Y)$$

nommée *inégalité de Schwartz*. En prenant la racine carrée des deux termes, on obtient

$$|cov(X,Y)| \leq \sigma(X).\sigma(Y)$$

et si $\sigma(X).\sigma(Y)$ est non nul, on peut diviser par $\sigma(X).\sigma(Y)$ d'où la relation cherchée $|\rho(X,Y)| \leq 1.\diamondsuit$

• *Démonstration de la propriété*

[R3]

Cet énoncé n'a aucun sens. La notion de variables aléatoires indépendantes n'a jamais été introduite. On consultera le chapitre 6 pour voir cette notion.

2.5 Démonstrations de certains calculs de m et σ pour les lois classiques

• *Calcul de m et σ pour la loi de Bernoulli*

Vérifions que les p_i définissent une probabilité. Puisque $0 \leq p \leq 1$, $p_1 = p$ et $p_0 = 1 - p$ sont des nombres positifs compris entre 0 et 1. De plus $p_0 + p_1 = (1-p) + p = 1.\diamondsuit$

$$\begin{array}{rcl} m_{b(p)} & = & \Sigma\ x_i.p_i \\ & = & 0.(1-p) + 1.p \\ & = & p \\ mt^2_{b(p)} & = & \Sigma\ x_i^2.p_i \\ & = & 0^2.(1-p) + 1^2.p \\ & = & p \\ V_{b(p)} & = & p - p^2 \\ & = & p(1-p)\diamondsuit \end{array}$$

• *Calcul de m et σ pour la loi Binomiale*

La méthode la plus simple consiste à utiliser la définition de $\mathcal{B}(n,p)$ comme somme de n variables de bernoulli indépendantes car alors

$$m_{\mathcal{B}(n,p)} = m(\ \Sigma\ b(p)\ ) = \Sigma\ m(\ b(p)\ ) = n.p$$

$$m_{\mathcal{B}(n,p)} = V(\ \Sigma\ b(p)\ ) = \Sigma\ V(\ b(p)\ ) = n.p.(1-p)$$

Mais il est également possible de calculer explicitement ces valeurs. Vérifions que les p_i définissent une probabilité. Puisque $0 \leq p \leq 1$, p^k et $(1-p)^{n-k}$ sont des nombres positifs compris entre 0 et 1. C_n^k est un entier positif et $\Sigma p_k = C_n^k.p^k.(1-p)^{n-k} = (p + (1-p))^n = 1$ d'après la formule du binome de Newton.$\diamondsuit$

$$
\begin{aligned}
m_{\mathcal{B}(n,p)} &= \sum_{k=0}^{n} x_k.p_k \\
&= \sum_{k=0}^{n} k.C_n^k.p^k.(1-p)^{n-k} \\
&= \sum_{k=1}^{n} (k.n!/k!(n-k)!).p^k.(1-p)^{n-k} \\
&= \sum_{k=1}^{n} n.((n-1)!/(\ (k-1)!((n-1)-(k-1))!\ ).p.p^{k-1}.(1-p)^{n-k} \\
&= n.p.\sum_{k=1}^{n} C_{n-1}^{k-1}.p^{k-1}.(1-p)^{n-k} = n.p.\sum_{k=0}^{n-1} C_{n-1}^{k}.p^k.(1-p)^{n-k} \\
&= n.p.[\ (p+(1-p)\ ]^{n-1} = n.p.1^{n-1} = n.p.
\end{aligned}
$$

Pour le moment d'ordre 2, on utilise la relation $k^2 = k(k-1)+k$:

$$
\begin{aligned}
mt^2_{\mathcal{B}(n,p)} &= \sum_{k=0}^{n} x_k^2.p_k \\
&= \sum_{k=0}^{n} k^2.C_n^k.p^k.(1-p)^{n-k} \\
&= \sum_{k=0}^{n} (\ k.(k-1)+k)\ ).C_n^k.p^k.(1-p)^{n-k} \\
&= n.(n-1).p^2.\sum_{k=2}^{n} .C_{n-2}^k.p^{k-2}.(1-p)^{n-k} + np \\
&= n.(n-1).p^2.1^{n-2} + n.p = n.(n-1).p^2 + n.p \\
V_{\mathcal{B}(n,p)} &= n.(n-1).p^2 + n.p - (n.p)^2 \\
&= n.p.(1-p)\diamondsuit
\end{aligned}
$$

• *Calcul de m et σ pour la loi Uniforme Discrète*

Vérifions que les p_k définissent une probabilité. Puisque $a < b$, les p_k qui valent tous $1/n$ sont des nombres positifs compris entre 0 et 1. De plus $\sum_{i=1}^{n} p_k = \sum_{i=1}^{n} 1/n = n.1/n = 1.\diamondsuit$

$$
\begin{aligned}
m_{\ \mathcal{U}(n)} &= \sum_{i=1}^{n} x_i.p_i &&= \sum_{i=1}^{n} i.\frac{1}{n} \\
&= \frac{n.(n+1)}{2}.\frac{1}{n} &&= (n+1)/2
\end{aligned}
$$

$$
\begin{aligned}
mt^2_{\mathcal{U}(n)} &= \sum_{i=1}^{n} x_i^2.p_i \quad = \sum_{i=1}^{n} i^2.\frac{1}{n} \\
&= \frac{n.(n+1).(2n+1)}{6}.\frac{1}{n} \quad = (n+1).(2n+1)/6 \\
V_{\mathcal{U}(n)} &= \frac{(n+1).(2n+1)}{6} - \left[\frac{n.(n+1)}{2}\right]^2 \\
&= \frac{(n+1)}{2}.\left[\frac{2n+1}{3} - \frac{n+1}{2}\right] \\
&= \frac{(n+1)}{2}.\left[\frac{4n+2-3n-3}{2}\right] \\
&= \frac{(n+1)}{2}.\left[\frac{n-1}{2}\right] \quad = (n^2-1)/12.
\end{aligned}
$$

• *Calcul de m et σ pour la loi Géométrique*

On utilise la fonction f définie en x par la somme de la série $\sum x^{k-1}$ pour $x < 1$. On démontre que $f(x) = 1/(1-x)$. On a également besoin de $f'(x) = 1/(1-x)^2$ et de $f''(x) = 2/(1-x)^3$.

Vérifions que les p_i définissent une probabilité. Puisque $0 \leq p \leq 1$, le nombre $p.(1-p)^{n-k}$ est positif et compris entre 0 et 1. De plus

$$\sum_{k \in \mathbb{N}} p.(1-p)^{k-1} = p.f(1-p) = p/(1-(1-p)) = p/p = 1/\diamondsuit$$

$$
\begin{aligned}
m_{\mathcal{G}(n,p)} &= \sum_{k \in \mathbb{N}} x_k.p_k \quad = \sum_{k \in \mathbb{N}} k.p.(1-p)^{k-1} \\
&= p.f'(1-p) \quad = p.1/p^2 \; = \; 1/p
\end{aligned}
$$

$$\begin{aligned}
mt^2{}_{\mathcal{G}(n,p)} &= \sum_{k\in\mathbb{N}} x_k^2.p_k \\
&= \sum_{k\in\mathbb{N}} k^2.p.(1-p)^{k-1} \\
&= \sum_{k\in\mathbb{N}} (\ k(k-1)+k\ ).p.(1-p)^{k-1} \\
&= p.(1-p).f''(1-p)+1/p \\
&= 2(1-p)/p^2+1/p \ = \ (2-p)/p^2 \\
V_{\mathcal{G}(n,p)} &= (2-p)/p^2-(1/p)^2 \ = \ (1-p)/p^2.
\end{aligned}$$

• *Calcul de* m *et* σ *pour la loi de Pascal*

Plutôt que d'effectuer des calculs, remarquons que la loi de Pascal peut être considérée comme la somme de s lois géométriques indépendantes. Car obtenir s boules peut être vu comme obtenir s fois 1 boule (*sic !*). D'où

$m_{\mathcal{P}}(n,p) = s/p$ et $V_{\mathcal{P}}(n,p) = s/(1-p)/p^2$.

• *Calcul de* m *et* σ *pour la loi de Poisson*

Vérifions que les p_i définissent une probabilité. Pour tout λ $e^{-\lambda}.\dfrac{\lambda^k}{k!}$ est un nombre positif. De plus, $\displaystyle\sum_{k\in\mathbb{N}} e^{-\lambda}.\frac{\lambda^k}{k!} = e^{-\lambda}.\sum_{k\in\mathbb{N}} \frac{\lambda^k}{k!} = e^{-\lambda}.e^{\lambda} = 1.\diamondsuit$

$$\begin{aligned}
m_{\ \mathcal{P}(\lambda)} &= \sum_{k\in\mathbb{N}} x_k.p_k \ = \ \sum_{k\in\mathbb{N}} k.e^{-\lambda}.\frac{\lambda^k}{k!} \\
&= e^{-\lambda}.\lambda \sum_{k\in\mathbb{N}} \frac{\lambda^{k-1}}{(k-1)!} \ = \ e^{-\lambda}.\lambda.e^{\lambda} \ = \ \lambda \\
{mt^2}_{\mathcal{P}(\lambda)} &= \sum_{k\in\mathbb{N}} x_k^2.p_k \ = \ \sum_{k\in\mathbb{N}} k^2.e^{-\lambda}.\frac{\lambda^k}{k!} \\
&= \sum_{k\in\mathbb{N}} (\ k(k-1)+k\ ).e^{-\lambda}.\frac{\lambda^k}{k!} \ = \ e^{-\lambda}.\lambda^2 \sum_{k\in\mathbb{N}} \frac{\lambda^{k-1}}{(k-1)!} \ + \lambda \\
&= \lambda^2+\lambda \\
V_{\ \mathcal{P}(\lambda)} &= \lambda^2+\lambda-(\lambda)^2 \ = \ \lambda
\end{aligned}$$

• *Calcul de m et σ pour la loi Uniforme Continue*

Vérifions que f est une densité de probabilité. f est constante sur $[a,b]$ et nulle ailleurs, soit : $f(x)=k$ si $x\in[a,b]$.

$$\begin{aligned}
\int_{\mathbb{R}} f &= \int_{-\infty}^{a} f + \int_a^b f + \int_b^{+\infty} f \\
&= \int_a^b k.dx \quad = k.[x]_a^b \quad = k(b-a)
\end{aligned}$$

Donc $\displaystyle\int f \ = 1$ pour $k(b-a)=1$ soit $k=1/(b-a)$.$\diamondsuit$

$$
\begin{aligned}
m_{\;\mathcal{U}([a,b])} &= \int_{\rm I\!R} Id.f = \int_a^b x.k.dx \\
&= k.\left[\frac{x^2}{2}\right]_a^b = (1/(b-a)).(b^2-a^2)/2 = (a+b)/2 \\
mt^2{}_{\mathcal{U}([a,b])} &= \int_{\rm I\!R} Id^2.f = \int_a^b x^2.k.dx \\
&= k.\left[\frac{x^3}{3}\right]_a^b = (1/(b-a)).(b^3-a^3)/2 = (a^2+ab+b^2)/3 \\
V_{\;\mathcal{U}([a,b])} &= (a^2+ab+b^2)/3 - ((a+b)/2))^2 \\
&= (4a^2+4ab+4b^2-3a^2-3b^2-6ab)/12 \\
&= (a^2-2ab+b^2)/12 = (a-b)^2/12.
\end{aligned}
$$

• *Calcul de* m *et* σ *pour la loi Loi Triangulaire*

f est un segment de droite sur $[a,m]$ pour $m = (a+b)/2$ qui vaut 0 en a. Donc $f(x) = \alpha.(x-a)$ et comme on impose $f(m) = 2/(b-a)$ on trouve $\alpha = 4/(b-a)^2$. Par le même raisonnement, $f(x) = -\alpha.(x-b)$ sur $[m,b]$. Montrons que f est une densité de probabilité.

$$
\begin{aligned}
\int_{\rm I\!R} f &= \int_a^b f = \int_a^m f + \int_m^b f \\
&= \int_a^m \alpha.(x-a)dx + \int_m^b -\alpha.(x-b)dx \\
&= \alpha.[(x-a)^2/2]_a^m - \alpha.[(x-b)^2/2]_m^b \\
&= \frac{\alpha\, a^2}{4} - \frac{\alpha\, ab}{2} + \frac{\alpha\, b^2}{4} \\
&= \alpha.(a^2-2ab+b^2)/4 = 1
\end{aligned}
$$

$$\begin{aligned}
m_{\mathcal{T}([a,b])} &= \int_{\mathbb{R}} Id.f = \int_a^m Id.f + \int_m^b Id.f \\
&= \alpha.\left[\frac{x(x-a)}{2}\right]_a^{(a+b)/2} - \alpha.\left[\frac{x(x-b)}{2}\right]_{(a+b)/2}^b \\
&= \frac{\alpha\ (2\,a+b)\,(a-b)^2}{24} + \frac{\alpha\ (a+2\,b)\,(a-b)^2}{24} \\
&= \frac{\alpha\ (a+b)\,(a-b)^2}{8} = (a+b)/2 \\
mt^2{}_{\mathcal{T}([a,b])} &= \int_{\mathbb{R}} Id^2.f = \int_a^m Id^2.f + \int_m^b Id^2.f \\
&= \alpha.\left[\frac{x^2(x-a)}{2}\right]_a^{(a+b)/2} - \alpha.\left[\frac{x^2(x-b)}{2}\right]_{(a+b)/2}^b \\
&= \frac{\alpha\ (11\,a^2+10\,ab+3\,b^2)\,(a-b)^2}{192} \\
&\quad + \frac{\alpha\ (3\,a^2+10\,ab+11\,b^2)\,(a-b)^2}{192} \\
&= \frac{\alpha\ (7\,a^2+10\,ab+7\,b^2)\,(a-b)^2}{96} \\
&= (7a^2+10ab+7b^2)/24 \\
V_{\mathcal{T}([a,b])} &= (7a^2+10ab+7b^2)/24 - ((a+b)/2))^2 \\
&= (a^2-2ab+b^2)/24 = (a-b)^2/24.
\end{aligned}$$

- *Calcul de m et σ pour la loi Exponentielle*

Vérifions que f est une densité.

$$\begin{aligned}
\int_{\mathbb{R}_+} f &= \int_0^{+\infty} \alpha.e^{-\alpha x}.dx \\
&= \alpha.\left[\frac{e^{-\alpha x}}{-\alpha}\right]_0^{+\infty} = -1.(0-1) = 1
\end{aligned}$$

$$\begin{aligned}
m_{\ \mathcal{E}(\alpha)} &= \int_{\mathbb{R}_+} Id.f = \int_0^{+\infty} x.\alpha.e^{-\alpha x}.dx \\
& \text{en intégrant par parties avec } u = x,\ v = e^{-\alpha x}/-\alpha \\
&= \alpha.\left[\frac{x.e^{-\alpha x}}{-\alpha}\right]_0^{+\infty} - \int_0^{+\infty} -1.e^{-\alpha x}.dx \\
&= 0\ - \left[\frac{e^{-\alpha x}}{-\alpha}\right]_0^{+\infty} = 1/\alpha \\
mt^2{}_{\mathcal{E}(\alpha)} &= \int_0^{+\infty} x^2.\alpha.e^{-\alpha x}.dx \\
& \text{en intégrant par parties avec } u = x^2,\ v = e^{-\alpha x}/-\alpha \\
&= \alpha.\left[\frac{x^2.e^{-\alpha x}}{-\alpha}\right]_0^{+\infty} - \int_0^{+\infty} -2.x.e^{-\alpha x}.dx \\
&= 0\ - 2.m/\alpha = 2/\alpha^2 \\
V_{\ \mathcal{E}(\alpha)} &= 2/\alpha^2 - (1/\alpha)^2 = 1/\alpha^2\ \diamondsuit
\end{aligned}$$

• *Calcul de m et σ pour la loi Gamma*

Vérifions que f est une densité.

$$\begin{aligned}
\int_{\mathbb{R}_+} f &= \int_0^{+\infty} \alpha^k.x^{k-1}.e^{-\alpha x}/\Gamma(k).dx \\
& \text{avec le changement de variable } y = \alpha x \\
&= \int_0^{+\infty} \alpha^k.(y/\alpha)^{k-1}.e^{-y}/\Gamma(k).d(y/\alpha) \\
&= \frac{\alpha^k}{\alpha^{k-1}.\Gamma(k)}.\int_0^{+\infty} y^{k-1}.e^{-y}.dy/\alpha \\
&= = \Gamma(k)/\Gamma(k) = 1
\end{aligned}$$

$$
\begin{aligned}
m_{\ \gamma(\alpha,k)} &= \int_{\mathbb{R}_+} Id.f = x.\alpha^k.x^{k-1}.e^{-\alpha x}/\Gamma(k).dx \\
& \quad \text{en intégrant par parties avec } u = x^k,\ dv = e^{-\alpha x} \\
&= \frac{\alpha^k}{\Gamma(k)} \left[\frac{e^{-\alpha x}}{-\alpha} x^k \right]_0^{+\infty} + \frac{k}{\alpha}. \int_{\mathbb{R}_+} f \\
&= 0 + (k/\alpha).1 = k\alpha
\end{aligned}
$$

$$
\begin{aligned}
mt^2{}_{\gamma(\alpha,k)} &= \int_{\mathbb{R}_+} Id^2.f = x^2.\alpha^k.x^{k-1}.e^{-\alpha x}/\Gamma(k).dx \\
& \quad \text{en intégrant par parties avec } u = x^{k+1},\ dv = e^{-\alpha x} \\
&= \frac{\alpha^k}{\Gamma(k)} \left[\frac{e^{-\alpha x}}{-\alpha} x^{k+1} \right]_0^{+\infty} + \frac{k+1}{\alpha}. \int_{\mathbb{R}_+} Id.f \\
&= 0 + ((k+1)/\alpha).m = k.(k+1)/\alpha^2 \\
V_{\ \gamma(\alpha,k)} &= k.(k+1)/\alpha^2 - (k/\alpha)^2 = k/\alpha^2
\end{aligned}
$$

- *Calcul de m et σ pour la loi de Weibull*

Vérifions que f est une densité.

$$
\begin{aligned}
\int_{\gamma}^{+\infty} f &= \int_{\gamma}^{+\infty} \frac{\beta}{\alpha}.\frac{x-\gamma}{\alpha}^{\beta-1} e^{-\frac{x-\gamma}{\alpha}^{\beta}}.dx \\
& \quad \text{avec le changement de variable } u = \tfrac{x-\gamma}{\alpha} \\
&= \int_{0}^{+\infty} \beta.u^{\beta-1} e^{-u^{\beta}} \\
& \quad \text{effectuons un second changement de variable } v = e^{-u^{\beta}} \\
&= \int_{0}^{+\infty} -d(\ e^{-v}\ ) \\
&= \left[-e^{-u^{\beta}} \right]_0^{+\infty} = -(1-0) = 1
\end{aligned}
$$

Nous aurons besoin du résultat suivant :

$$
\begin{aligned}
I_n &= \int_0^{+\infty} u^n.\beta.e^{-u^\beta}.du \text{ avec le changement de variable } v = u^\beta \\
&= \int_0^{+\infty} v^{n/beta}.e^{-v}.dv = \Gamma(\frac{n}{\beta}+1)
\end{aligned}
$$

$$
\begin{aligned}
m_{\mathcal{W}(\alpha,\beta,\gamma)} &= \int_0^{+\infty} Id.f = \int_0^{+\infty} x.\frac{\beta}{\alpha}.\frac{x-\gamma}{\alpha}^{\beta-1}.e^{-\frac{x-\gamma}{\alpha}^\beta}.dx \\
&\quad \text{avec le changement de variable } u = \tfrac{x-\gamma}{\alpha} \\
&= \int_0^{+\infty} (\alpha.u+\gamma).\beta.u^{\beta-1}.e^{-u^\beta}.du \\
&= \alpha.I_1 + \gamma.1 = \alpha.\Gamma(1+1/\beta)+\gamma
\end{aligned}
$$

$$
\begin{aligned}
mt^2{}_{\mathcal{W}(\alpha,\beta,\gamma)} &= \int_0^{+\infty} Id^2.f = x^2.\frac{\beta}{\alpha}.\frac{x-\gamma}{\alpha}^{\beta-1} e^{-\frac{x-\gamma}{\alpha}^\beta}.dx \\
&= \int_0^{+\infty} (\alpha.u+\gamma)^2.\beta.u^{\beta-1}.e^{-u^\beta}.du \\
&= \alpha^2.I_2 + 2.\alpha.\gamma.I_1 + \gamma^2.1 \\
&= \alpha^2.\Gamma(2+1/\beta) + 2.\alpha.\gamma.\Gamma(1+1/\beta) + \gamma^2 \\
V_{\mathcal{W}(\alpha,\beta,\gamma)} &= \alpha^2.\Gamma(1+1/\beta) + 2.\alpha.\gamma.\Gamma(1+1/\beta) + \gamma^2 - (\alpha.\Gamma(1+1/\beta)+\gamma)^2 \\
&= \alpha.^2(\ \Gamma(1+2/\beta) - \Gamma(1+1/\beta)^2\ ).
\end{aligned}
$$

• *Calcul de m et σ pour la loi Normale*

Commençons par établir quelques résultats annexes.

$$I = \int_0^{+\infty}\int_0^{+\infty} e^{-(x^2+y^2)}.dx.dy$$

avec le changement de variable $x = \rho.cos(\theta)$, $y = \rho.sin(\theta)$

$$= \int_0^{+\pi/2}\int_0^{+\infty} e^{-\rho^2}\rho.d\rho.d\theta$$

$$= \int_0^{+\pi/2} d\theta.\int_0^{+\infty} \rho.e^{-\rho^2}.d\rho$$

$$= [\theta]_0^{\theta/2}.\left[\frac{-1}{2}e^{-r^2}\right]_0^{+\infty}$$

$$= (\pi/2).(\ 0-(-1/2)\ ) \ = \ \pi/4$$

$$J = \int_0^{+\infty}\int_0^{+\infty} x^2y^2e^{-(x^2+y^2)}.dx.dy$$

avec le changement de variable $x = \rho.cos(\theta)$, $y = \rho.sin(\theta)$

$$= \int_0^{+\pi/2} sin^2(\theta).cos^2(\theta).d\theta.\int_0^{+\infty} \rho^5.e^{-\rho^2}.d\rho$$

effectuons un second changement de variable $u = \rho^2$

$$= \int_0^{+\pi/2} \frac{1-cos(4\theta)}{8}.d\theta.\frac{1}{2}\int_0^{+\infty} u^2.e^{-u^2}.du$$

$$= \frac{1}{8}\left[\theta - \frac{sin(4\theta)}{4}\right].\frac{1}{2}\Gamma(3)$$

$$= (\pi/16).2/2 \ = \ \pi/16$$

Montrons que f est bien une densité.

$$I = \int_{-\infty}^{+\infty} \frac{1}{\sigma.\sqrt{2\pi}}.e^{-(\frac{x-\mu}{\sigma})^2/2}.dx$$

avec le changement de variable $u = (x-\mu)/\sigma$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}.e^{-u^2/2}.du$$

effectuant un second changement de variable $v = u/\sqrt{2}$

$$= \frac{1}{\sqrt{\pi}}.2\int_{0}^{+\infty} e^{-v^2}.dv$$

$$= \frac{2}{\sqrt{\pi}}.\sqrt{I} = 1$$

$$m_{\mathcal{N}(m,\sigma)} = \int_{-\infty}^{+\infty} Id.f = \int_{-\infty}^{+\infty} \frac{1}{\sigma.\sqrt{2\pi}}.x.e^{-(\frac{x-\mu}{\sigma})^2/2}.dx$$

avec le changement de variable $u = (x-\mu)/\sigma$

$$= \frac{\sigma}{\sqrt{\pi}}.\int_{-\infty}^{+\infty} u.e^{-u^2/2}.du + \mu.\int_{-\infty}^{+\infty} f$$

$$= 0 + \mu.1 = \mu$$

$$
\begin{aligned}
{mt^2}_{\mathcal{N}(m,\sigma)} &= \int_{-\infty}^{+\infty} Id^2.f = \int_{-\infty}^{+\infty} \frac{1}{\sigma.\sqrt{2\pi}}.x^2.e^{-(\frac{x-\mu}{\sigma})^2/2}.dx \\
& \quad \text{avec les mêmes changements de variable qu'au-dessus} \\
&= \frac{\sigma^2}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} u^2.e^{-u^2/2}.du \ + \frac{2\sigma\mu}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} u.e^{-u^2/2}.du \ + \mu^2.\int_{-\infty}^{+\infty} f \\
&= \frac{\sigma^2}{\sqrt{2\pi}}.2\int_{-\infty}^{+\infty} u^2.e^{-u^2/2}.du \ + 0 + \mu^2.1 \\
&= \frac{\sigma^2}{\sqrt{\pi}}.4.\int_{0}^{+\infty} v^2.e^{-v^2}.dv \ + \mu^2 \\
&= \frac{\sigma^2}{\sqrt{\pi}}.4.\sqrt{J} + \mu^2 \ = \ \sigma^2 + \mu^2 \\
V_{\mathcal{N}(m,\sigma)} &= \sigma^2 + \mu^2 - (\mu)^2 \ = \ \sigma^2
\end{aligned}
$$

Chapitre 3.

Statistique

3.1 Statistiques Descriptives

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 une telle partie de cours et nous nous bornerons dans les exercices à sensibiliser à ces problèmes. On trouvera par contre dans la bibliographie les références nécessaires aux ouvrages correspondants.

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*.

3.1.1 Variables et séries

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 de valeurs ordonnées. 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 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 variable non renseignée ou mal renseignée est source d'erreur. On prendra donc soin à connaitre le sens de chaque code, chaque unité.

3.1.2 Variables qualitatives et quantitatives

Nous restreignons volontairement pour l'instant nos données à deux types élémentaires, les *variables à unités* et les *variables à codes* nommées aussi variables quantitatives et variables qualitatives. Le chapitre 4 viendra présenter les traitements associés à d'autres types de variables, tout aussi élémentaires quant à leur origine mais plus délicates à manipuler comme les variables multi-réponses, les variables textuelles... Nous allons supposer aussi que toutes nos variables sont transcrites numériquement, laissant à l'informaticien le problème du codage, recodage, formatage 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 n'on en fait aucune importance, parce qu'elles correspondent à un codage arbitraire. 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. Cela se fait souvent à l'aide de fichiers externes nommés descriptifs. 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.

3.1.3 Analyse univariée

On dit aussi analyse à une seule dimension ou analyse séparée pour bien indiquer qu'on prend en compte chaque colonne de valeurs indépendamment des autres colonnes. L'analyse univariée d'un dossier de n variables est donc équivalent à l'analyse des n dossiers constitués chacun d'une seule variable. Pour une variable quantitative, on effectue un calcul calqué sur le celui des variables aléatoires : moyenne, variance, écart-type etc. Pour une variable qualitative, on effectue (et le nom peut sembler curieux pour l'instant) 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) $p_j = 100 * fj$.

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 des modalités (on parle alors de disproportion).

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).

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 moyenne puis un deuxième par ordre décroissant de coefficient de variation σ/m.

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 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. On doublera très souvent les 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").

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. En particulier, on se méfiera de l'équirépartition, base de nombreux sondages qui peut être criticable ; ainsi, pour un dossier comme ELF, la disproportion apparente entre hommes et femmes est justifiée par le thème de l'enquête. 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 ? C'est là tout un art que de "noyer la 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.

Nous livrons ici quelques adages qui devraient être le *credo* de tout statisticien, même de l'apprenti et qu'il faut relire avant chaque analyse à effectuer, après chaque rédaction :

- La machine calcule, l'humain interpréte 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.

3.1.4 Analyse bivariée

Encore nommée analyse à deux dimensions, on vient ici s'intéresser aux couples de variables. A priori, pour n variables en tout, on en a C_n^2 à traiter, soit $n.(n-1)/2$ mais souvent on n'effectue pas tous ces traitements. Tout d'abord parce qu'il y en a trop à effectuer, ensuite par ce qu'ils n'ont pas toujours de sens. On vient donc effectuer tous les croisements entre variables quantitatives, qu'on nomme corrélations et on effectue seulement certains croisements nommés tris croisés pour les variables qualitatives. Il arrive qu'on vienne aussi faire des traitements par sous-populations, utilisant les modalités de variable qualitative comme critère de découpage, mais on retombe alors sur un problème d'analyse univariée sur un sous-ensemble de la population.

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. Il y a mathématiquement corrélation si $|\rho|$ est égal à 1. En pratique, 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 $d = m(X)^2 - mt(X,2)$

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). Nous avons détaillé ces pièges dans les exercices.

Pour des variables qualitatives ordinales, la notion de corrélation peut avoir un sens (corrélation au sens de Kendall, de Spearman), mais c'est dans un contexte bien particulier que nous traiterons plus lois. 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.

3.2 Comparaison et Approximation

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 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 cherche en général un seul indicateur global, ce qui oblige souvent à utiliser une expression à base de sommesde différences.

Soit en effet à comparer deux répartitions a_i et b_i pour un même effectif global Nt. On mettra en oeuvre des expressions comme $\sum(a_i - b_i)$, $\sum (a_i - b_i)^2$, $\sum |a_i - b_i|$ etc. La sous-section suivante est consacrée à la caractérisation de telles valeurs.

3.2.1 Dissimilarités, Distances...

Soit E un ensemble et d une application de ExE dans R^+. On ne suppose rien de plus sur E. Ce pourra donc être un ensemble de nombres, un ensemble de fonctions, un ensemble de matrices, d'espaces vectoriels... La notion de distance, de dissimilarité, d'écart... est liée à trois notions : une notion de symétrie qui permettra de dire "x et y sont proches" sans distinguer de sens ; une notion métrique qui permettra de quantifier la "taille" de la distance ; une notion de "dissemblance" ou de dissimilarité qui permet d'être sûr que les éléments, quand ils sont différents, sont à une certaine "distance" les uns des autres. On conçoit donc que la notion de distance est une notion générale, "sympathique", applicable à des domaines concrets et à des structures complexes aussi bien en mathématiques pures qu'en mathématiques appliquées, en statistiques, en qualité, en physique...

On dit que d est un indice de dissimilarité sur E ou plus simplement une dissimilarité ssi elle vérifie la condition :

$$(\forall x,y) \quad d(x,y) = 0 \Leftrightarrow x = y$$

On rappelle que d est dite symétrique ssi

$$(\forall x,y) \quad d(x,y) = d(y,x)$$

De plus, d est dite sous-maximale ssi

$$(\forall x,y,z) \quad d(x,y) \leq max(d(x,z),d(z,y))$$

Enfin, d est dite triangulaire ssi

$$(\forall x,y,z) \quad d(x,y) \leq d(x,z) + d(z,y)$$

Une distance est une dissimilarité symétrique et triangulaire. Une ultramétrique est une dissimilarité symétrique sous-maximale. Un ensemble muni d'une distance est appelé espace métrique. Il existe de nombreuses distances dont on trouvera la définition dans n'importe quel ouvrage d'analyse. La notion de distance permet de définir des notions de topologie générale (boule ouverte, ensemble ouvert, filtre...) et de convergence (continuité, compacité, complétude, suite de Cauchy...). On montre qu' une ultramétrique est une distance.

Soit a_i, b_i, c_i trois répartitions de nombres strictement positifs de même somme Nt. On nomme coefficient *khi-deux* (noté χ^2) des répartitions a_i et b_i de centre c_i la quantité

$$\chi^2_c(a,b) = \sum (a_i - b_i)^2 / c_i$$

Si les c_i sont fixés, l'application $(a,b) \rightarrow \sqrt{\chi^2_c(a,b)}$ définit une fonction qui est un indice de dissimilarité symétrique et triangulaire. C'est donc une distance.

Deux problèmes se posent ensuite : quelles valeurs prendre pour les c_i ? comment décider que deux a_i et b_i sont proches si leur χ^2 est non nul ? A la première question, il est facile de répondre si on chercher à comparer a_i à une répartition théorique b_i : on prendra $c_i = b_i$. Sinon, on calculera une répartition théorique à partir du tableau de contingence (ou tri croisé) induit par les a_i et les b_i. Pour la deuxième, on utilise une "table" de χ^2.

Reste encore à savoir quelle loi théorique utiliser pour b_i. Admettant que les a_i correspondent à des valeurs i =0,1,2...n on peut utiliser deux modèles classiques : la loi Binomiale et la loi de Poisson tronquée. Dans chacun des cas, il faut déterminer les paramètres de ces lois, calculer les probabilités théoriques, passer des probabilités aux fréquences puis à des effectifs entiers avant de calculer le chi-deux. Nous prendrons ici des exemples pour montrer ces calculs.

A partir de t valeurs i et a_i, on calcule les fréquences observées $obs_i = a_i/Nt$ où Nt est l'effectif total $\sum a_i$. La moyenne observée est donc $m_{obs} = \sum i.obs_i$. Si les a_i correspondent à une loi Binomiale $B(n,p)$ de moyenne théorique m_{th} alors on doit avoir $n = t$ et $m_{obs} = m_{th}$ ce qui permet de déduire $p = m_{obs}/n$. Connaissant n et p, on dresse la liste des probabilités théoriques p_i associées aux valeurs i, i de 0 à n. La fréquence théorique est donc $th_i = p_i.Nt$ et le chi-deux calculé est alors $\chi^2_{ht}(obs,th)$.

3.2.2 Calculs concrets de χ^2

Comme premier exemple de calcul de χ^2, considérons le résultat du tirage de 200 boules parmi 4 couleurs : rouge, vert, blanc et bleu. Il y a une proportion de 30 % de boules rouges, 20 % de boules vertes, 40 % de boules blanches et 10 % de boules bleues. On fournit le tirage suivant : 70 boules rouges, 40 vertes, 68 blanches et 22 bleues. Pour déterminer le χ^2, on commmence par déterminer le nombre théorique de boules par couleur. Il sufit pour cela de multiplier les proportions par le nombre total de boules, ce qui nous donne 200*30 % = 60 boules rouges, 40 vertes, 80 blanches et et 20 bleues. Le χ^2 est donc $(60-70)^2/60 + (40-40)^2/40 + (80-68)^2/80 + (20-22)^2/20$ soit 3.667.

Comme deuxième exemple, on fournit les résultats d'une expérience où on a effectué plusieurs lancers de 6 pièces de monnaie. On a obtenu 18 lancers "0 ou 1 fois pile", 42 lancers "2 fois pile", 66 lancers "3 fois pile", 45 lancers "4 fois pile" et 21 lancers "5 ou 6 fois pile". Le nombre total de lancers est donc 192. La loi théorique est la loi binomiale $B(6{,}0.5)$ à condition de regrouper les valeurs 0 et 1 d'un coté, les valeurs 5 et 6 de l'autre. Il est possible de calculer par programme informatique ou avec une calculatrice les probabilités p("X=0") = 0.0156 = p("X=6"), p("X=1") = 0.0938 = p("X=5"), p("X=2") = 0.2344 = p("X=4") et p("X=3")=0.3125. Le nombre théorique de "0 ou 1 pile" est donc de (0.0156+0.0938)*192 soit 21.0048 arrondi en 21 mais ce n'est pas la méthode la plus pratique ici. En effet, pour la loi binomiale, ici p=1-p donc chaque probabilité est proportionelle au C_n^p correspondant. On a donc une fréquence de 1/64 pour "X=0" et "X=6", 6/64 pour "X=1" et "X=5", 15/64 pour "X=2" et "X=4", 20/64 pour "X=3". Puisque 192=3*64, le nombre de "0 ou 1 fois pile" est donc (1+6)*3 soit 21, valeur exacte. De même, on a 45 lancers "2 fois", 60 lancers "3 fois", 45 lancers "4 fois", 21 "5 ou 6 fois pile". Le χ^2 est donc $(18-21)^2/21 + (42-45)^2/45 + (66-60)^2/60 + (45-45)^2/20 + (21-21)^2/21$ soit 1.260.

3.2.3 Approximations par la loi Binomiale et la loi de Poisson

Des essais récents sur 145 échantillons de 30 cartes-mères de P8 fournissent le tableau suivant de nombre de cartes-mères défectueuses :

x_i	0	1	2	3	4	5	6	7	...	30
n_i	44	55	29	10	5	1	1	0	...	0

On vérifie que la somme totale des effectifs $Nt = \sum n_i$ vaut bien 145. La loi binomiale à considérer est ici $B(,np)$ avec $n = 30$. Pour p, puisque la moyenne pondérée vaut $m = \sum x_i.n_i/Nt = 1.2 = n.p$, on déduit que $p = 0.04$. Les 12 premières valeurs de la loi binomiale $B(30{,}0.04)$ ainsi que leur produit par 145 sont

x_i	p_i	$p_i.145$	*arrondi*
0	0.2938576	42.6093520	43
1	0.3673221	53.2617045	53
2	0.2219237	32.1789365	32
3	0.0863037	12.5140365	13
4	0.0242729	3.5195705	4
5	0.0052591	0.7625695	1
6	0.0009130	0.1323850	0
7	0.0001304	0.0189080	0
8	0.0000156	0.0022620	0
9	0.0000016	0.0002320	0
10	0.0000001	0.0000145	0
11	0.0000000	0.0000000	0

Les termes d'ordre supérieur à 12 sont négligeables. La somme des 9 premières probabilités est 0.9999825, soit à près multiplication par 145, un total de 144.9974625. Pour appliquer le calcul du χ^2, il faut regrouper des classes. On obtient finalement les valeurs théoriques 43, 53, 32 et 17 mises en regard des valeurs observées 44, 55, 29 et 17. Le χ^2 est alors de 14.

3.2.4 χ^2 d'indépendance et χ^2 d'ajustement

Il arrive qu'on veuille comparer p "caractères" ou variables qualitatives selon la répartition qu'elles induisent sur une même population. On parle alors de χ^2 d'indépendance alors que les χ^2 précédents étaient nommés χ^2 d'ajustement à une loi théorique. Pour un tel calcul, la valeur théorique correspond à l'indépendance des caractères qui est exprimée par le fait qu'on peut calculer cette valeur théorique en effectuant le produit du total de la ligne par le total de la colonne divisé par le total général.

Pour le χ^2 d'indépendance, les notations classiques sont les suivantes. Soient $n_{i,j}$ les valeurs du tri croisé. On note $n_{i.}$ le total de la ligne i et $n_{j.}$ le total de la colonne j. Soit $n_{..}$ le total général, qui vaut donc $\sum_j \; n_{j.}$ mais aussi $\sum_i \; n_{i.}$. Le χ^2 est alors calculé pour tous les $obs_k = n_{i,j}$ et pour tous les $th_k = n_{i.} \; n_{.j}/n_{..}$ où on parcourt toutes les valeurs de i et j par exemple en prenant k de 1 à $l * c$ si l est le nombre de lignes et c le nombre de colonnes avec $k(i,j) = (c-1) * i + j$. Le nombre de degrés de liberté est alors $(l-1).(c-1)$.

```
6 lignes et 4 colonnes

   1      0      0      0  +     1
  11      4      3      6  +    24
  11      1      7      0  +    19
   6      4     11      4  +    25
   7     12     11     10  +    40
   1      1      1      3  +     6
+++++++++++++++++++++++++++++++++
  37     22     33     23  +   115

   0.322      0.191      0.287      0.200   +      1.000
   7.722      4.591      6.887      4.800   +     24.000
   6.113      3.635      5.452      3.800   +     19.000
   8.043      4.783      7.174      5.000   +     25.000
  12.870      7.652     11.478      8.000   +     40.000
   1.930      1.148      1.722      1.200   +      6.000
+++++++++++++++++++++++++++++++++++++++++++++++++++++
  37.000     22.000     33.000     23.000   +      0.000

 chi-deux 28.151
```

A titre d'exemple, voici un programme Awk qui calcule ce χ^2 d'indépendance

```
{ for (i=1;i<=NF;i++) { d[FNR,i]=$i ; tg += $i }
  nbl = FNR ; nbc = NF }

END{ print nbl " lignes et " nbc " colonnes  "  ;

 d[nbl+1,nbc+1] = tg # total general

 for (i=1;i<=nbl;i++) { for (j=1;j<=nbc;j++) {
   tl[i] += d[i,j] } ; d[i,nbc+1] = tl[i] } # total ligne

 for (j=1;j<=nbc;j++) { for (i=1;i<=nbl;i++) {
   tc[j] += d[i,j] } ; d[nbl+1,j]   = tc[j] } # total colonne

 for (i=1;i<=nbl+1;i++) { # affichage avec les marges
     for (j=1;j<=nbc+1;j++) { printf(" %4d ",d[i,j]) ;
        if (j==nbc) {printf(" + ")} }
        print " "
        if (i==nbl) { print("++++++++++++++++++++") }  }
 print " "

 for (i=1;i<=nbl;i++) {  # valeurs theoriques
  for (j=1;j<=nbc;j++) { t[i,j] = tl[i]*tc[j]/tg ;
           x = (t[i,j]-d[i,j]); ch += x*x/t[i,j] }   }

 for (i=1;i<=nbl;i++) { for (j=1;j<=nbc;j++) { # total ligne
   hl[i] += t[i,j] } ; t[i,nbc+1] = hl[i] }

 for (j=1;j<=nbc;j++) { for (i=1;i<=nbl;i++) { # total colonne
   hc[j] += t[i,j] } ; t[nbl+1,j]   = hc[j] }

 for (i=1;i<=nbl+1;i++) {
     for (j=1;j<=nbc+1;j++) {
        printf(" %8.3f ",t[i,j])
        if (j==nbc) {printf(" + ")} }
        print " "
        if (i==nbl) { print("+++++++++++++++++++++") }  }
 print " "
 print " chi-deux " ch
}
```

Pour se ramener à un coefficient positif entre 0 et 1, on utilise diverses techniques de normalisation de ce χ^2 d'indépendance. Notant χ pour le chi-deux précédent et N le total général, ces coefficients sont

$$\begin{aligned}
\text{C de Pearson} &= \sqrt{\frac{\chi}{\chi+N}} \\
\text{T de Tschuprow} &= \sqrt{\frac{\chi}{N.\sqrt{(l-1)(c-1)}}} \\
\text{V de Cramer} &= \sqrt{\frac{\chi}{N.inf\{l-1,c-1\}}} \\
\text{G de vraisemblance} &= 2\sum_i \sum_j n_{ij}.ln(N.n_{i.}/(n_{i.}n_{.j}))
\end{aligned}$$

La contribution d'une case en ligne i, colonne j au chi-deux vaut $n_{..}(n_{ij} - n_{i.}n_{.j})^2/(\chi n_{i.}n_{.j})$. Une valeur élevée d'une telle contribution met en évidence une association significative entre la modalité i et la modalité j. Signalons enfin un dernier indice, mais non symétrique, le τ_b de Goodman et Kruskal :

$$\tau_b(y/x) \ = \ \frac{(\sum_i \sum_j n_{ij}^2/(N.n_{i.}) - T}{1-T}$$

avec $T \ = \ 1 - \sum_j (n_{.j}/N)^2$

3.2.5 Corrélation des rangs

L'analyse bivarée de variables ordinales (ou rangs, ou variables hiérarchiques) a donné lieu à de nombreux indices de corrélation. Considérant les rangs comme des variables quantitatives, Spearman propose de calculer le classique coefficient de corrélation. Mais comme les données sont des permutations de $\{1,2...n\}$ on obtient, en posant $\varepsilon_i = a_i - b_i$ (différence entre deux rangs) la quantité

$$\rho_S = 1 - 6.\sum_i \varepsilon_i^2/D$$

avec $D = n.(n^2-1)$ où n est le nombre de termes à classer.

Le coefficient de Kendall se calcule en comptant pour chaque a_i le nombre de b_j supérieurs à a_i pour $j > i$ (les a_i étant ordonnés). Si R est la somme de ces nombres alors

$$\rho_K = \frac{4.R}{n(n-1)} - 1$$

Sur un exemple numérique, on saisit mieux la qualité de ces coefficients. Soient les deux classements sur 10 objets

a_i	1	2	3	4	5	6	7	8	9	10
b_i	3	1	2	4	6	5	9	8	10	7

alors n = 10, $\sum_i (a_i - b_i)^2$ = 22, D = 990, ρ_S = 1 -6*22/990 soit 0.87. De même, R = (8+7+7+6+4+4+2+1+0) = 39 donc ρ_K = (4*39/90) - 1 = 0.73

3.2.6 Comparaison de deux pourcentages

Soient A et B deux échantillons de taille n_a et n_b contenant respectivement une proportion p_a et p_b d'individus marqués. Pour savoir si ces deux proportions correspondent à deux échantillons provenant d'une même population avec une proportion p d'individus marqués, on effectue le calcul de "l'écart réduit" défini par

$$\varepsilon = |p_a - p_b|/r \text{ où } r = \sqrt{p(1-p)(1/n_a + 1/n_b)}$$

si p est la proportion d'individus marqués dans la population regroupant n_a et n_b. La différence est dite significative au risque de $\alpha = 5$ % si ε est supérieur à 1.96 ; cela signifie qu'on a seulement 5 chances sur 100 de se tromper en affirmant que les proportions ne viennent pas d'une même population.

Par exemple, sur 102 ordinateurs parisiens achetés, 20 sont des bi-Pentium ; sur 98 ordinateurs angevins achetés, 28 sont des bi-Pentium. On a donc n_a = 102, n_b = 98, p_a = 20/102 = 0.196078431, p_b = 28/98 = 0.285714286, p = (20+28)/(102+98) = 0.24 , $\varepsilon = 0.089635855/\sqrt{0.00364945978} = 1.48377303$... donc on peut estimer que ces deux proportions sont issues d'une mêeme population. Par contre, avec 10 fois plus de personnes (soit 1020 ordinateurs parisiens et 980 angevins), on aurait un écart réduit de $1.48 * \sqrt{10}$ soit 4.69 et la conclusion serait inversée.

3.2.7 Comparaison de deux moyennes

Soient A et B deux échantillons de taille n_a et n_b de moyenne respective m_a et m_b et de variance respective v_a et v_b. Pour savoir si ces deux moyennes correspondent à deux échantillons provenant d'une même population, on effectue un calcul similaire à celui de "l'écart réduit" défini pour la comparaison de pourcentages. Soit

$$\delta = |m_a - m_b|/r \text{ où } r = \sqrt{v_a/n_a + v_b/n_b)}$$

La différence est dite significative au risque de $\alpha = 5$ % si δ est supérieur à 1.96 ; cela signifie qu'on a seulement 5 chances sur 100 de se tromper en affirmant que les moyennes ne viennent pas d'une même population.

Amusons-nous (!) à comparer des notes d'élèves issus de France (série a_i) et de Navarre (série b_i). On donne dans le tableau suivant les effectifs par note (pour des notes de 0 à 10, ce qui montre que cet exemple ne correspond pas à "la" réalité) :

note	a_i	b_i
0	0	3
1	0	3
2	3	6
3	9	11
4	7	7
5	5	13
6	17	10
7	10	12
8	11	4
9	4	1
10	1	3

Puisque $n_a = 67$, m_a=5.85074627, v_a= 3.97772332, $n_b = 73$, m_b=4.87671233, v_b= 5.64233440, $\delta = 0.9740/\sqrt{0.136661257}$ soit 2.63482262 donc la différence est significative.

3.3 Introduction à l'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, 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. 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. 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 n 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 qualitatives ou d'un 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 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.

A titre d'exemple, nous présenterons les données les résultats du dossier *ANTAL* dont nous avons déjà parlé. Nous commençons par le descriptif du fichier :

`Antal.Dsc`

```
  Pour  comparer  l'effet  de  deux  médicaments antalgiques
  (d'oú  le  nom   ANTAL  pour  le   dossier)  et  de   leur
  association, on soumet à  40 patients ces traitements  "en
```

```
double insu", c'est-à-dire que ni le patient ni le docteur
ne savent de quel traitement il s'agit.  Les quatre
traitements sont notés A, B, & et P. & est en fait
l'association de A et B, P est un placebo (absence de
médicament).
```

```
On note dans échelle de temps repérée par les
identificateurs de lignes la réaction des patients.  Ainsi
A1/4 signifie après 1/4 d'heure pour le traitement A, P2h
signifie après 2 heures pour le placebo etc.  Les colonnes
signifient respectivement douleur nulle (DOU0), douleur
légère (DOU1), douleur modérée (DOU2), sévère (DOU3), très
sévère (DOU4) ; on indique aussi une douleur théorique
(DOU5) qui devrait être associée voire équivalente à DOU4.
```

Les données fournies sont :

```
     HEUR  DOU0 DOU1 DOU2 DOU3 DOU4 DOU5
     A0h      0    0    0   28   12   25
     A1/4     0    2    4   28    6   13
     A1/2     0    4   16   18    2    5
     A1h      6    4   18   10    2    5
     A2h      6   10   12    8    4    9
     A3h      6   10    8   12    4    9
     A4h      4   12    4   16    4    9
     A>5h     0   14    6   16    4    9
     A>6h     0   14    6   16    4    9
     BP0h     0    0    0   22   18   37
     B1/4     0    4    6   22    8   17
     B1/2     0   12   10   16    2    5
     B1h      0   14   16    8    2    5
...
     &0h      0    0    0   18   22   45
     &1/4     0    4    6   20   10   21
     &1/2     0    4   20   14    2    5
     &1h      4   12   18    6    0    1
     &2h      4   20   12    4    0    1
     &3h     12   14    7    7    0    1
     &>3h    12   10   10    8    0    1
     &>4h    12   10   10    8    0    1
     &>5h    12   10   10    8    0    1
```

```
     P0h      0    0    0   22   18   37
     P1/4     0    0    0   22   18   37
     P1/2     0    0    0   34    6   13
     P1h      2    4   16   14    4    9
     P2h      2    8    8   10    4    9
...
```

Une ASG (Analyse Statistique Générale) fournit alors les résultats suivants :

```
 ETUDE de la base  ANTAL
     Il y a          36  lignes dans votre base
         et          07  colonnes
 Nødu   Nom du   Moyenne  Ecart-Type  Coef. Var. Min     Max        Etendue
 champ  champ       m         s        s/m
   1 HEUR      non numérique ; son type est :  Caractère
   2 DOU0         2.944     3.756     1.276     0.000     12.000     12.000
   6 DOU4         5.444     5.899     1.084     0.000     22.000     22.000
   7 DOU5        11.889    11.799     0.992     1.000     45.000     44.000
   3 DOU1         8.278     5.541     0.669     0.000     20.000     20.000
   4 DOU2         8.639     5.583     0.646     0.000     20.000     20.000
   5 DOU3        14.472     6.784     0.469     4.000     34.000     30.000

 Matrice des corrélations de  ANTAL  avec         36  lignes
        DOU0    DOU1    DOU2    DOU3    DOU4    DOU5

DOU0    1.0000
DOU1    0.3344  1.0000
DOU2    0.2229  0.4055  1.0000
DOU3   -0.5996 -0.7055 -0.6716  1.0000
DOU4   -0.4627 -0.7227 -0.6925  0.5548  1.0000
DOU5   -0.4627 -0.7227 -0.6925  0.5548  1.0000  1.0000


 Borne de corrélation sûre     ( 0.9 conseillé )          0.9
 Borne de corrélation possible ( 0.6 conseillé )          0.6
 Meilleures corrélations
 *  DOU5  DOU4      1.000
    DOU4  DOU1     -0.723
    DOU5  DOU1     -0.723
    DOU3  DOU1     -0.705
    DOU4  DOU2     -0.693
    DOU5  DOU2     -0.693
    DOU3  DOU2     -0.672
```

```
Formules linéaires associées
   Corrélation    1.000 : DOU5   =    2.000 * DOU4   +    1.000
   Corrélation    1.000 : DOU4   =    0.500 * DOU5   +   -0.500
```

L'AFC du fichier produit les résultats suivants :

```
*****************************************
* B I B L I O T H E Q U E   A D D A D *
*        MICRO   (VERSION 89.1)         *
*        17/09/89                       *
*****************************************


     ANALYSE DES CORRESPONDANCES (ANCORR)
     D'APRES : YAGOLNITZER ET TABET


VALEURS  PROPRES       VAL(1)=   1.00000
--------------------------------------------------------------------
NUM ! VALPRO. ! POURC.! CUMUL !VARIAT.!*! HISTO. DES VALEURS PROPRES
---------------------------------------------------------------------
2 !  .29516 ! 42.430! 42.430!*******!*!***************!***************
3 !  .21473 ! 30.868! 73.298! 11.562!*!***************!*******
4 !  .12113 ! 17.412! 90.710! 13.456!*!************
5 !  .06457 !  9.283! 99.993!  8.130!*!*******
6 !  .00005 !   .007!100.000!  9.275!*!

    ! I1 ! QLT POID INR!  1#F  COR CTR!  2#F  COR CTR!  3#F  COR CTR!
--------------------------------------------------------------------
   1!A0h !1000   17  37! 1046  734  65! -502  170  21! -344   80  17!
   2!A1/4!1000   35  14!  419  622  21!  156   87   4! -204  147  12!
   3!A1/2!1000   35  11!   -3    0   0! -157  108   4! -394  683  45!
   4!A1h !1000   35  37! -668  600  53!   19    0   0!  545  399  86!
   5!A2h !1000   35  31!  381  239  17! -328  177  17!  595  582 102!
   6!A3h !1000   35  37! -260   93   8!   49    3   0!  741  751 158!
   7!A4h !1000   35   6!  115  121   2!  -93   78   1!  296  797  25!
   8!A>5h!1000   35   7!   -1    0   0!  126  112   3! -300  640  26!
   9!A>6h!1000   35   7!   -1    0   0!  126  112   3! -300  640  26!
  10!BP0h!1000   17  65! 1062  435  67! 1190  546 115!  221   19   7!
...
  35!P5h !1000   35  34! -772  869  71!  237   82   9! -150   33   6!
  36!P6h !1000   17  36! -903  571  48!  -28    1   0! -146   15   3!

    ! J1 ! QLT POID INR!  1#F  COR CTR!  2#F  COR CTR!  3#F  COR CTR!
--------------------------------------------------------------------
   1!DOU0!1000    0   0!  900   70   0!-1868  303   0!   12    0   0!
   2!DOU1!1000  115 166! -252   63  25! -418  174  94!  848  717 685!
   3!DOU2!1000  189 154! -535  505 183!   89   14   7! -362  230 204!
   4!DOU3!1000  229 165! -596  710 276!   88   15   8!  -16    1   1!
   5!DOU4!1000  281 243!  565  531 305! -495  407 321! -185   57  80!
   6!DOU5!1000  185 271!  580  331 212!  813  649 570!  142   20  31!
--------------------------------------------------------------------
    !    !        1000!         1000!         1000!         1000!
```

```
AXE HORIZONTAL( 1)--AXE VERTICAL( 2)--TITRE:AFC ANTAL.DBF
 NOMBRE DE POINTS :   42 ==ECHELLE :  4 CARACTERE(S) = .070    1 LIGNE = .029
 +-------------------------------------------------------+-----------------------------------------------------------BP0h+  2  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 ...
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                   DOU5                        !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       B1/4                                                            !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                           P1/2!  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                       P2h                             !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !           P5h                                         !                                                               !  0  0  1
 !                                                       !   P4h                                                         !  0  0  1
 &1/4                                        P3h         !                                                               !  0  0  1
 !                                               P1h     !                       A1/4                                    !  0  0  1
 !                                                       A>5h                                                            !  1  0  1
 !                       DOU3DOU2                        !                                                               !  0  0  1
 !                                   B3h     A3h         !                                                               !  0  0  1
 !                   A1h                                 !                                                               !  0  0  1
 +-------------------------------------------------------+---------------&1/2--------------------------------------------+  0  0  1
 !   B4h P6h                                             !                                                               !  0  0  1
 !                       B6h                             !                                                               !  0  0  1
 !                                                       !       A4h                                                     !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       A1/2                                                            !  1  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                       B1/2                    !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                       A2h                                     !  0  0  1
 !                                           &3h         !                                                               !  0  0  1
 !                                                       !       B2h                                                     !  0  0  1
 !                                       &1h DOU1        !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                               DOU4                        A0h !  1  0  1
 !                                                   B5h !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !   &2h                                                         !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                               !  0  0  1
 !                                                       !                                           &>3h                !  2  0  1
 !                                                       !                                                               !  0  0  1
 ...
 !                                                       !                                                               !  0  0  1
 !                                                       !                                                   DOU0        !  0  0  1
 +-------------------------------------------------------+---------------------------------------------------------------+  0  0  1
 NOMBRE DE POINTS SUPERPOSES :   7
 P0h (BP0h)  P1/4(BP0h)  A>6h(A>5h)  B1h (A1/2)  &0h (A0h )  &>4h(&>3h)  &>5h(&>3h)
```

Une suite logique de l'AFC est une CAH soit les fichiers :

```
CLASSIFICATION ASCENDANTE HIERARCHIQUE (CAH2CO),  METHODE DES VOISINS REDUCTIBLES

SOMME DES INDICES DE NIVEAU    .69565E+00
--------------------------------------------------------------------------------
!  J  ! I(J) ! A(J)! B(J)!T(J)!T(Q)! HISTOGRAMME DES INDICES DE NIVEAU
--------------------------------------------------------------------------------
!   71!   170!   70!   54! 244! 244!****************************************
!   70!   140!   63!   69! 201! 445!*********************************
!   69!    86!   68!   66! 123! 568!********************
!   68!    66!   67!   62!  94! 663!***************
!   67!    41!   65!   61!  59! 721!**********
!   66!    33!   58!   56!  47! 768!********
!   65!    30!   60!   64!  43! 811!*******
...
!   46!     1!   14!   17!   2! 997!*
!   45!     1!   31!   33!   2! 999!*
!   44!     1!   32!   35!   1!1000!*
!   43!     0!    1!   19!   0!1000!*
!   42!     0!    3!   13!   0!1000!*
!   41!     0!   27!   40!   0!1000!*
!   40!     0!   25!   26!   0!1000!*
!   39!     0!   29!   38!   0!1000!*
!   38!     0!   10!   28!   0!1000!*
!   37!     0!    8!    9!   0!1000!*

--------------------------------------------------------------------------
!  J  ! I(J) ! A(J)! B(J)! P(J)! CLASSES DE LA HIERARCHIE
----------------------------------------------------------
!   71!   170!   70!   54!   36!
----------------------------------------------------------
!   70!   140!   63!   69!   32! A0h  &0h  &>5h &>3h &>4h
!     !      !     !     !     ! A1/4 A1/2 B1h  A>5h A>6h
!     !      !     !     !     ! B1/2 &1/2 B6h  &1/4 P2h
!     !      !     !     !     ! A3h  B3h  A2h  A4h  ...
----------------------------------------------------------
!   69!    86!   68!   66!   27! P4h  P1h  P3h  B1/4 A1/4
!     !      !     !     !     ! &1h  &3h  B2h  B5h  B1/2
!     !      !     !     !     ! P5h  P6h  A1h  B4h  A3h
...
!   39!     0!   29!   38!    3! P1/4 BP0h P0h
--------------------------------------------------------
!   38!     0!   10!   28!    2! BP0h P0h
--------------------------------------------------------
!   37!     0!    8!    9!    2! A>5h A>6h
--------------------------------------------------------
```

REPRESENTATION DE LA CLASSIFICATION HIERARCHIQUE

```
     +--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+


A0h  *----*---------------------------------------------------------------------------*----------------*-
     !    !                                                                           !                !
&0h  -    !                                                                           !                !
          !                                                                           !                !
&>5h *-----                                                                           !                !
     !                                                                                !                !
&>3h !                                                                                !                !
     !                                                                                !                !
&>4h -                                                                                !                !
                                                                                      !                !
P4h  -*--*-----------*------*-------------*-----------*--------------------------------                !
      !  !           !      !             !           !                                                !
P1h  *-  !           !      !             !           !                                                !
     !   !           !      !             !           !                                                !
P3h  -   !           !      !             !           !                                                !
         !           !      !             !           !                                                !
B1/4 -----           !      !             !           !                                                !
                     !      !             !           !                                                !
A1/4 --*--------*-----      !             !           !                                                !
       !        !           !             !           !                                                !
A1/2 *--        !           !             !           !                                                !
     !          !           !             !           !                                                !
B1h  -          !           !             !           !                                                !
                !           !             !           !                                                !
A>5h !          !           !             !           !                                                !
     !          !           !             !           !                                                !
A>6h -          !           !             !           !                                                !
                !           !             !           !                                                !
&1h  *-----------           !             !           !                                                !
     !                      !             !           !                                                !
&3h  -                      !             !           !                                                !
                            !             !           !                                                !
B2h  *---*-------------------             !           !                                                !
     !   !                                !           !                                                !
B5h  -   !                                !           !                                                !
         !                                !           !                                                !
B1/2 *----                                !           !                                                !
     !                                    !           !                                                !
&1/2 -                                    !           !                                                !
                                          !           !                                                !
B6h  -*---*--------------------------------           !                                                !
      !   !                                           !                                                !
&1/4 --   !                                           !                                                !
          !                                           !                                                !
P2h  *---*-                                           !                                                !
     !   !                                            !                                                !
P5h  -   !                                            !                                                !
         !                                            !                                                !
P6h  -----                                            !                                                !
                                                      !                                                !
A1h  -*-*--------------*-------------------------------                                                !
      ! !              !                                                                               !
B4h  -- !              !                                                                               !
        !              !                                                                               !
A3h  --*-              !                                                                               !
       !               !                                                                               !
B3h  ---               !                                                                               !
                       !                                                                               !
A2h  -**----------------                                                                               !
      !!                                                                                               !
A4h  --!                                                                                               !
       !                                                                                               !
&2h  ---                                                                                               !
                                                                                                       !
P1/4 **-------------------------------------------------------------------------------------------------
     !!
BP0h !!
     !!
P0h  -!
      !
P1/2 --
```

3.4 Démonstrations

Chapitre 4.

Autres variables et Traitements

4.1 Variables numériques moins classiques

Un premier cas de variables 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	....	- du thé	....
- du lait	....	- du jeu d'orange	....

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 de variables 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 : [1]

> "classer de 1 à 3 ces chateaux de 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.

Une variable **QF** ou *Question Floue* est une variable à modalitée pour laquelle on attribue une valeur probabiliste (nombre réel entre 0 et 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.

1. où est le piège ?

4.2 Questions ouvertes et Variables textuelles

Les "vraies" questions ouvertes sont ensuite beaucoup plus délicates à mettre en oeuvre et nous les incluons dans les **QX** ou *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"

Une QX peut être une réponse à une question ouverte, un texte complet, un récit, un compte-rendu d'expertise...[2] Traiter une QX est d'un tout autre ordre que traiter une QT ou une QL. 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. Nous emploierons le terme "élément de texte" pour regrouper les deux vocables forme graphique et lemme. En mode non lemmatisé, "chanteur", "chanteuse" et "chanteuses" compteront 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.

On s'intéresse alors autant aux termes les plus fréquents (mais certains sont

2. 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 (comme pour une simple question en oui/non), il faut aussi accorder une importance à l'absence de réponse pour une QX.

prévisibles[3]) 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. Il reste ensuite à essayer de décrire ce que cela semble induire.

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...

4.3 Traitements Statistiques évolués

Les traitements statistiques que nous avons présentés jusqu'ici, sauf peut-être celui des variables textuelles, étaient assez élémentaires dans la mesure où ils mettaient en jeu des données faciles à manipuler. Mais on peut aller plus loin. De nombreux domaines industriels mettent en jeu des scènes, des représentations graphiques, comme par exemple les images en télé-detection satellite, ou en télé-surveillance. Les données élémentaires ne sont plus ici des nombres mais des images. Une analyse d'images est une analyse statistique particulière. En fonction du domaine, on viendra essayer de trouver des caractéristiques synthétiques comme les couleurs, les formes, les contours. C'est encore de la statistique mais à un niveau plus élevé.

3. Lesquels ?

De même, il est possible de s'intéresser aux séquences animées qui ne sont jamais que des suites (ou vecteurs) d'images. La difficulté rajoutée est ici la dimension temporelle. Les sons peuvent aussi être captés, numérisés, que ce soit les bruits de l'univers en astrophysique ou des sons aquatiques en biologie sous-marine ou même des partitions musicales. Depuis quelques années, les sémiologues s'intéressent aux dynamiques gestuelles. Les données sont alors des listes de gestes, des attitudes... Dans un autre domaine, en chimie, biochimie, bactériologie, il n'est pas rare que les données élémentaires soient des séquences d'ADN, des liste des composants chimiques, des spectres...

Il est clair que le terme statistique peut s'appliquer à tous ces domaines. Les problèmes posés sont ceux de la reconnaissance des types de données. Il est peu clair qu'on puisse ramener une image à une QL ou à une QT. Analyser une bande-son n'est pas visiblement la même chose que de traiter une QH... C'est pourquoi toutes ces analyses procèdent de l'Analyse des Données plutôt que des Statistiques. Une fois résolu le problème du codage, de la transcription des données, on se trouve en présence de tableau de données de grandes dimensions.

Nous ne traiterons bien sûr aucun de ces tableaux de données. Mais nous pouvons donner quelques gardes-fous. Tout d'abord, il importe de cerner le but de l'étude statistique. Suivant que l'on veut caractériser ou que l'on veut comparer, à partir du moment où on veut classer (mettre dans des classes pré-établies) ou classifier (construire des classes), la démarche n'est pas la même. Si on cherche à décrire, les outils statistiques ne sont pas les mêmes que si on cherche à prévoir. Si on cherche à expliquer, on n'utilise pas les mêmes fonctions que si on veut prédire...[4]

Il importe beaucoup d'essayer de dégager les composants élémentaires en vue de l'analyse statistique. Reprenons l'exemple des images en télé-détection. Le stockage informatique d'une image est souvent fait (independamment de la technique de stockage, qui peut mettre en jeu des techniques de compression sophistiquées) en termes de pixels, c'est à dire de points élémentaires. Pour une image radar, les points ne sont pas la bonne structure élémentaire. C'est la forme, le contour qui l'est. Il faut donc effectuer un premier traitement pour dégager les composantes du contour et stocker les paramètres de ce contour avant de traiter le fichier de l'ensemble des contours.

4. Quelle est la différence entre prévoir et prédire ?

L'Analyse des Données est donc une analyse statistique évoluée.Elle requiert des outils mathématiques globaux plus algébriques que la statistique classique (en particulier de l'algèbre linéaire). Elle demande aussi une bonne culture générale en statistique et informatique pour adapter des méthodes ou en concevoir de nouvelles...

4.4 Exercices sur ces variables et traitements

- *QM incomplètes*
  Dans la présentation des QM au début de ce chapitre, les questions sur le petit déjeuner sont mal posées. Pourquoi ? Que manque-t-il aux énoncés ?
- *Télévision, Culture et Choix de question*
  Critiquez le texte de la question sur la télévision qui sert de présentation aux variables QP.
- *Total des choix pour une QP*
  Après traitement d'une QP, quel est le total des moyennes des pourcentages ? Justifiez votre réponse.
- *Scores et Anti-scores*
  Quelle différence de traitement doit-on faire suivant qu'on a des scores ou des anti-scores ?
- *Choix de l'échelle des valeurs*
  Comment décider si on utilise un score, une note sur 10, sur 20, un pourcentage ?
- *Questionnaire jeu-de-piste*
  Que penser d'un questionnaire où on demande à la question 8. de répondre par oui ou par non puis où on indique : "si vous avez répondu oui, allez à la question 12 sinon allez à la question 15 ?"
- *Initiales*
  Dans le cadre d'une enquête sur des patients cérébro-lésés pour un hopital, il était demandé de fournir des noms d'animaux, de légumes, de vêtements puis de donner des noms communs commençant par P, L ou T. D'autres auteurs, notamment américains semblent préférer des mots commençant par L, T ou M ? Quel choix est "raisonnable" et pourquoi ?
- *La loi de Zipf*
  Qu'est-ce que la loi de Zipf en Statistique Lexicale ?
- *Analyse textuelle*
  Si on effectue l'analyse d'un texte français de plusieurs pages, quels éléments seront les plus fréquents ?
- *Statistiques lexicales internationales*
  Peut-on reconnaitre une langue par la longueur de ses mots, de ses phrases, par les fréquences des lettres initiales de mots ?
- *Statistique lexicale et Cryptographie*

Quand on essaie de percer un message secret codé, un problème souvent considéré comme élémentaire est celui de savoir dans quelle langue il est écrit. Quelle méthode (rapide) proposez-vous pour faire trouver à un ordinateur dans quelle langue un texte est écrit ?

Chapitre 5.

Vecteurs et Convergence

Ce chapitre est un un peu plus difficile que les autres car il met en jeu des notions mathématiques plus techniques. On commence par définir la notion de fonction caractéristique (qui n'est jamais qu'une transformée de Fourier) avant de passer aux vecteurs aléatoires dont les vecteurs gaussiens sont l'utilisation la plus fréquente. On passe ensuite en revue les différents modes de convergence et on conclut par des approximations classiques de lois et les fameuses "lois des grands nombres". Même si certains calculs sont délicats ou peu clairs, la philosophie sous-jacente et les conclusions sont très importantes en pratique, puisqu'elles permettent de considérer les probabilités comme limites de fréquences. Les probabilités et les statistiques sont donc encore plus interdépendantes qu'il n'y paraissait jusqu'ici.

Le lecteur ou la lectrice peu familié(e) avec l'intégration, les espaces vectoriels, est invité(e) à parcourir les annexes pour ne pas être perdu(e) dans l'exposé.

5.1 Indépendance et Convolution

Nous avions défini dans le chapitre 2 la notion d'indépendance pour deux deux événements A et B par la propriété : $p(A \cap B) = p(A).p(B) \Leftrightarrow A$ et B sont dits *indépendants*. Allons un peu plus loin. On dira que deux variables aléatoires X et Y définies sur le même espace probabilisé sont indépendantes ssi $p(\text{"}X \in A_i\text{"} \cap \text{"}Y \in B_j\text{"}) = p(\text{"}X \in A_i\text{"}).p(\text{"}Y \in B_j\text{"})$ pour tout couple de boréliens A_i et B_j, ce qu'on pourra donc noter $p_{(X,Y)} = p_X.p_Y$.

Si X et Y admettent respectivement les densités f et g, alors (X,Y) admet comme densité $f.g$. De plus, $m(XY)$ est alors $m(X).m(Y)$.

La convolution des variables aléatoires X et Y est la variable aléatoire somme $Z = X + Y$. On s'intéresse alors surtout aux variables indépendantes. Dans le cas discret, si X et Y sont deux telles variables aléatoires, alors on a directement $p("Z = z") = \sum_x p("X = x").p("Y = z - x")$ ce qui permet par exemple de montrer que la somme de deux variables de Poisson indépendantes de paramètres λ_1 et λ_2 est encore une loi de Poisson de paramètre $\lambda_1 + \lambda_2$. Dans le cas continu, la densité h de Z est définie par la convolution mathématique classique, à savoir $k(z) = \int f(u)g(z - u)du$ où f est la densité de X et g celle de Y.

Montrons le résultat annoncé pour les lois de Poisson. Soit $X = \mathcal{P}(\lambda_1)$ et $Y = \mathcal{P}(\lambda_2)$. Si $Z = X + Y$ alors

$$\begin{array}{rcl} p("Z = z") & = & \sum_x p("X = x").p("Y = z - x") \\ & = & \sum_x \; e^{-\lambda_1}\lambda_1^x/x! \; . \; e^{-\lambda_2}\lambda_2^{z-x}/(z-x)! \\ & = & e^{-(\lambda_1+\lambda_2)} \; . \; \sum_x \lambda_1^x \lambda_2^{z-x}/(x!(z-x)!) \\ & = & e^{-(\lambda_1+\lambda_2)} \; . \; \sum_x C_z^x . \lambda_1^x \lambda_2^{z-x}/z! \\ & = & e^{-(\lambda_1+\lambda_2)} \; . \; (\lambda_1 + \lambda_2)^z/z! \; \diamondsuit \end{array}$$

On peut de même montrer que la convolution de deux loix uniformes continues sur [0,1] est une fonction dont la densité est $1 - |x - 1|$ si x est dans [0,2] et 0 ailleurs. Nous laissons en exercice la discussion sur la somme de deux variables aléatoires binomiales indépendantes $\mathcal{B}(n_1,p_1)$ et $\mathcal{B}(n_2,p_2)$.

Il serait par contre plus délicat et maladroit de vouloir démontrer que la somme de deux lois normales indépendantes est encore une loi normale. Il existe un outil beaucoup plus puissant pour cela : les fonctions caractéristiques, que nous présentons dans la section suivante.

5.2 Fonctions caractéristiques

On apelle *fonction caractéristique* d'une variable aléatoire X la transformée de Fourier de la probabilité p_X. On la note traditionnellement φ ou φ_X s'il faut préciser la variable aléatoire mise en jeu. On a donc par définition

$$\varphi_X(t) = m(\ e^{itX}\ ) = \int_{\mathbb{R}} e^{itx} dP_X(x)$$

Si X est discrète alors

$$\varphi_X(t) = \sum_k e^{itx_k} p_k$$

et si X possède une densité f alors

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f(x).dx$$

Il est intéressant de remarquer que

[K1] $\varphi_{\lambda X}(t) = \varphi_X(\lambda t)$
[K2] $\varphi_{X+a}(t) = e^{ita}\varphi_X(t)$
[K3] $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$ si X et Y sont indépendantes
[K4] $\varphi_X^{(k)}(0) = i^k m(X^k)$

On peut s'amuser (!) à démontrer que

[L1] $\varphi_{b(p)}(t) = p.e^{it} + (1-p)$
[L2] $\varphi_{\mathcal{B}(n,p)}(t) = (\varphi_{b(p)}(t))^n$
[L3] $\varphi_{\mathcal{P}(\lambda)}(t) = e^{\lambda.(e^{it}-1)}$
[L4] $\varphi_{\mathcal{U}([-a,a])}(t) = sin(at)/at$
[L5] $\varphi_{\mathcal{N}(m,\sigma)}(t) = e^{itm}e^{-t^2\sigma^2/2}$

Outre le calcul des moments de X, φ sert a démontrer simplement des propriétés pour la convergence des suites de variables aléatoires et simplifie dans de nombreux cas les calculs pour des variables indépendantes. On aura l'occasion de s'en rendre compte dans la section suivante. L'un des problèmes posés par la notion de fonction caractéristique est le niveau élevé des calculs qui fait appel à l'Analyse Complexe et notamment au calcul des résidus. C'est pourtant une notion fondamentale.

Il est possible de montrer que φ_X existe toujours (car p_X est une mesure bornée) et qu'elle est toujours continue. Un résultat important est le suivant : deux variables aléatoires ayant même fonction caractéristique ont même loi de probabilité. De plus, moyennant certaines conditions, à partir d'une fonction φ, on peut trouver X telle que $f = \varphi_X$ grâce à la formule d'inversion de la transformée de Fourier. En particulier pour montrer que X est la loi normale $\mathcal{N}(0,1)$, il suffit de montrer que $ln(\varphi_X(t)) = -t^2/2$. Il est donc simple de montrer que la somme de deux lois normales indépendantes $\mathcal{N}(\mu_1,\sigma_1)$ et $\mathcal{N}(\mu_2,\sigma_2)$ est encore une loi normale $\mathcal{N}(\mu,\sigma)$ avec $\mu = \mu_1 + \mu_2$ et $\sigma = \sqrt{\sigma_1 + \sigma_2}$:

$$\begin{aligned}
\varphi_{\mathcal{N}(\mu_1,\sigma_1)+\mathcal{N}(\mu_2,\sigma_2)}(t) &= e^{it\mu_1}e^{-t^2\sigma_1^2/2} \;.\; e^{it\mu_2}e^{-t^2\sigma_2^2/2} \\
&= e^{it(\mu_1+\mu_2)} \;.\; e^{-t^2(\sigma_1+\sigma_2^2/2} \\
&= \varphi_{\mathcal{N}(\mu_1+\mu_2,\sqrt{\sigma_1+\sigma_2}\,)}(t)
\end{aligned}$$

5.3 Convergence

5.3.1 Les différents modes de convergence

Si on dispose d'une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ et d'une variable aléatoire X définies sur un même espace probabilisé $(\Omega,\mathcal{T},p)$ il y a plusieurs façons d'envisager la convergence des X_n vers X.

Convergence en moyenne d'ordre k

Si les moments d'ordre k de la variable $X_n - X$ existent, on dit que les X_n convergent en moyenne d'ordre k vers X *ssi* la suite numérique réelle des $m(|X_n - X|^p)$ converge vers 0.

Convergence en probabilité

Les X_n convergent en probabilité vers X *ssi*

$$\forall \varepsilon,\eta\ \exists N\ n > N \ \Rightarrow p(|X_n - X| \geq \varepsilon) \leq \eta$$

Convergence presque sûre

Deux variables aléatoires X et Y sont égales presque sûrement *ssi*

$$p(\{\omega | X(\omega) \neq Y(\omega)\}) = 0$$

La suite $(X_n)_{n\in\mathbb{N}}$ converge presque sûrement vers X *ssi*

$$p(\{\omega | \lim_{n\to+\infty} X_n(\omega) \neq X(\omega)\}) = 0$$

Convergence en loi

Les X_n convergent en loi vers X *ssi* en tout point de continuité u de la fonction de répartition F de X, la suite $F_n(u)$ des fonctions de répartition des X_n en u converge vers $F(u)$.

Tous ces modes de convergence ne sont pas indépendants. On peut démontrer les résultats suivants :

[C1]	La convergence d'ordre k	implique	la convergence en probabilité
[C2]	La convergence presque sûre	implique	la convergence en probabilité
[C3]	La convergence en probabilité	implique	la convergence en loi

Nous conviendrons d'utiliser les notations suivantes

$X_n \xrightarrow{m^k} X$	pour (X_n) converge à l'ordre k	vers X
$X_n \xrightarrow{ps} X$	pour (X_n) converge presque sûrement	vers X
$X_n \xrightarrow{p} X$	pour (X_n) converge en probabilité	vers X
$X_n \xrightarrow{\mathcal{L}} X$	pour (X_n) converge en loi	vers X

5.3.2 Quelques convergences remarquables

[T1] *Théorème de Moivre-Laplace*
Si (X_n) est une suite de $\mathcal{B}(n,p)$ indépendantes,
alors $(X_n - np)/\sqrt{np(1-p)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

[T2] Si (X_λ) est une suite de $\mathcal{P}(\lambda)$ indépendantes,
alors $(X_\lambda - \lambda)/\sqrt{\lambda} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ pour $\lambda \to +\infty$

[T3] *Théorème "central-limite"*
Si (X_n) est une suite de variables aléatoires indépendantes et de même loi, de même moyenne μ et de même écart-type σ
alors $(\Sigma X_i - n\mu)/(\sigma\sqrt{n}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

[T4] *Loi faible des grands nombres*
Si (X_i) est une suite de variables aléatoires indépendantes,
de moyennes respectives finies m_i et d'écart-types respectifs σ_i finis,
alors, pour $(\Sigma m_i)/n \to m$ et $(\Sigma\sigma_i^2)/n^2 \to 0$ on a $(\Sigma X_i)/n \xrightarrow{p} m$

[T5] *Loi forte des grands nombres*
Si (X_n) est une suite de variables aléatoires indépendantes telles que
$(\Sigma m_i)/n \to m$ et $(\Sigma\sigma_i^2)/n^2$ est convergente alors $(\Sigma X_i)/n \xrightarrow{ps} m$

Pour démontrer ces deux derniers théorèmes, nous avons besoin des inégalités suivantes

[I1] *Inégalité de Markov*
Si X est une variable aléatoire réelle, positive et intégrable,
alors $p("X \geq r") \leq m(X)/r$

[I2] *Inégalité de Markov d'ordre alpha*
Si α est un nombre positif et X une variable aléatoire réelle
telle que $m(|X^\alpha|)$ est finie,
alors $p("X \geq r") \leq m(|X|^\alpha)/r^\alpha$

[I3] *Inégalité de Tchebyscheff*
Si X est une variable aléatoire réelle de moment d'ordre 2 fini,
alors pour tout r on a $p("|X - m(X)| \geq r") \leq \sigma^2(|X|)/r^2$

[I4] *Inégalité des grands nombres*
Si les (X_i) sont des variables aléatoires réelle d'ordre 2,
de même loi et indépendantes,
alors pour $S_n = \displaystyle\sum_{i=1}^{n} X_i$ on a $p("|S_n/n - m| \geq \varepsilon") \leq \sigma^2(|X|)/(n.\varepsilon^2)$
où m est la moyenne commune et σ l'écart-type commun.

[I5] *Loi faible de Bernoulli*
Soit $(A_n)_{n \in \mathbb{N}}$ une suite d'évènements de même probabilité p
et deux à deux indépendants. Si N_n est le nombre de réalisations des A_k
pour $k \leq n$ et $\varphi_n = N_n/n$ la fréquence de réalisation
des A_k pour $k \leq n$ alors pour tout $\varepsilon > 0$ on a
$p("|\varphi_n - p| \geq \varepsilon") \leq 1/4n\varepsilon^2$.

5.4 Vecteurs Aléatoires Gaussiens

5.4.1 Vecteurs Aléatoires

Un vecteur aléatoire X est une application de $(\Omega,\mathcal{T},p)$ dans un espace vectoriel E. En pratique, nous utiliserons $E = \mathbb{R}^n$ ou $E = \mathbb{C}^n$. X est donc un vecteur $(X_1,X_2...X_n)$ de variables aléatoires. La moyenne de X est alors le vecteur des moyennes de chaque composante de X. On appelle *matrice de variance-covariance* de X la matrice notée Σ (ou Σ_X si on a besoin de préciser le vecteur aléatoire) définie par $\Sigma_{i,j} = cov(X_i,X_j)$. Avec des notations matricielles, $\Sigma_X = m(X \odot {}^tX) - m(X) \odot {}^tm(X)$ où ${}^tV$ désigne la transposée du vecteur V et où $\odot$ est le produit cartésien induit par le produit scalaire. Si les X_i sont des variables centrées réduites, Σ_X est alors la matrice des corrélations de X notée $M_\rho(X)$. Si M est une matrice de dimensions telles que $Y = M.X$ a un sens, alors

[M1] $m(MX) = M.m(X)$
[M2] $\Sigma_{MX} = M.\Sigma_X.{}^tM$

Le théorème suivant, dit de caractérisation des matrices de variance-covariance a d'importantes conséquences pour l'Analyse des Données :

[CMV] Toute matrice symétrique positive M est
la matrice de variance-covariance d'un vecteur aléatoire.

La fonction caractéristique du vecteur aléatoire X est définie en $a \in E$ par $\varphi_X(a) = e^{ia.tX}$ où $V.W$ est le produit scalaire usuel. Le résultat le plus important sur les fonctions caractéristiques des vecteurs aléatoires est le théorème d'indépendance :

[THI] Les X_i sont indépendantes *ssi* la fonction caractéristique de X
est égale au produit des fonctions caratéristiques des X_i.

Par contre, pratiquement, on utilise le théorème de Cramer-Wold :

[CW] La loi de X est entièrement déterminée par celles
des combinaisons linéaires des X_i.

5.4.2 Vecteurs Gaussiens

On dit que X est un vecteur gaussien si toute combinaison linéaire de ses composantes est une loi normale. Si A est une matrice et B un vecteur, il est facile de montrer que $A.X+B$ est gaussien.

Soit X un vecteur gaussien centré. Alors, pour $a \in E$:

[CC] $\varphi_X(a) = e^{-a.^t\Sigma_X.a/2}$

On en déduit que

[IG] Pour un vecteur gaussien X, les X_i sont indépendantes *ssi* Σ_X est une matrice diagonale.

Chapitre 6.

Statistique et Informatique

S'il était possible dans les années 50 d' enseigner les probabilités et les statistiques avec un crayon et du papier seulement, il faut aujourd'hui savoir utiliser un logiciel statistique, avoir des idées pour (re)programmer les calculs usuels. Ce chapitre passe donc en revue ce qu'il faut savoir pour aller des statistiques livresques aux statistiques sur machine. On commence par y décrire la représentation et le stockage des données, puis on présentera quelques logiciels pour finir par des exemples d'algorithmes et de programmes qui mettent en pratique des notions vues dans les chapitres précédents.

6.1 Fichiers et 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, comme le format .DBF de Dbase, le format .XLS d'Excel. 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). 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 et oblige à travailler avec la longueur maximale des chaines.

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 a l'avance toutes les plages de variation, il faut très souvent se soucier de l'ordre de grandeur des données pour prévoir l'affichage.

Si le stockage d'une variable quantitative 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 une variable qualitative, il est souvent suffisant de conserver les codes de modalités. Pour une variable a deux modalités codées 1 et 2 la liste des 1 et des 2 est claire, à conditionque 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 10 et 01. 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 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). Pour les codages flous pondérés (*cf.* le chapitre précédent, 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 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. Pour des tableurs, un coup d'oeil au cadrage droit ou gauche[2] permet parfois de détecter tout de suite pourquoi le moindre calcul statistique aboutit inévitablement à un message du genre *Invalid Data Type*.

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...

2. Pourquoi ?

6.2 Logiciels et Programmation

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, 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 !

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 la 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

```
 MODULE AnaQt
  Globales   X        /* tableau des donnÅes  */
             n        /* nombre de valeurs    */
  Locales    k        /* indice de boucle     */
             sv       /* somme des valeurs    */
             sc       /* somme des carrÅs     */
  Fonction   externe :  rac  /* racine carrÅe  */
 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

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

avec une erreur inférieure à 10^-7 pour

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

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
     paramÉtre lambda           2.8
     nombre de termes           7
       xi  puissance factorielle probabilitÅ
         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...

ANNEXES