Probabilités, statistiques : une introduction

Erwan Penchèvre

2013–2014

1 Evénements et aléas

Qu'est-ce qu'un *modèle* mathématique ? C'est une description d'un ou plusieurs phénomènes c'est-à-dire une certaine configuration dans le temps et dans l'espace. Cette description se fait souvent au moyen de variables quantitatives, c'est-à-dire de quantités pouvant prendre plusieurs valeurs possibles suivant les différentes modalités du phénomène modélisé. Ce qui importe est alors de formuler les relations entre ces variables. On essaie, si possible, d'isoler certains paramètres et des variables libres dont les autres variables dépendent entièrement par des relations fonctionnelles.

Exemple n° 1 : idéalement, un modèle climatique permet de prévoir la température en tout lieu terrestre (ϕ, θ, z) à la date t. Les prévisions dépendent de certains paramètres, par exemple la quantité de gaz à effet de serre.

Exemple n° 2 : je jette un dé à six faces. Un modèle idéal prédirait de manière certaine la trajectoire du dé sur la table, l'instant où le dé s'arrête de rouler, et la face supérieure visible à cet instant. La trajectoire dépend de certains paramètres : l'impulsion initiale que je donne au dé en le lançant, le caractère plus ou moins lisse de la table (le dé va peut-être glisser), etc.

On a quelque espoir de réussir à concevoir des modèles climatiques, au moins à échelle réduite et courte durée, mais lorsqu'on lance un dé, on se dit qu'il est impossible de prédire le résultat. C'est le principe même des jeux de hasard. Cela n'interdit pas d'utiliser les mathématiques pour modéliser de tels phénomènes ; cependant, la valeur de certaines variables ne sera pas entièrement déterminée par les variables libres et les paramètres. On parle alors de *variable aléatoire* ou *aléa*. La théorie des probabilités consiste en l'étude de telles variables.

Une autre notion importante en probabilités est celle d'*événement*. Le possible n'est pas unique, il existe en général plusieurs possibles, et les probabilités ne cherchent pas à donner une description d'un réel déterminé de manière certaine mais plutôt une description de tous les possibles, assortie d'une certaine mesure du caractère plus ou moins probable de chacun de ces possibles. On est donc conduit à découper la réalité étudiée en *événements*, chaque événement étant une configuration spatio-temporelle possible lorsque le phénomène se réalise. On note les événements avec des lettres majuscules (A, B, C, ...).

Une *mesure de probabilité* est une fonction qui associe à chaque événement A un nombre $p(A)$ compris entre zéro et un (parfois exprimé comme pourcentage) :

$$p : A \longmapsto p(A) \in [0, 1]$$

Remarque : $p(A) = 0$ si l'on juge l'événement A très improbable voire impossible. On a $p(A) = 1$ si l'on juge au contraire cet événement certain (100 %) ; alors on dit que A se produit *presque sûrement*.

Définition 1.1 *Un* espace de probabilité $(\Omega, \mathfrak{S}, p)$ *est un modèle décrivant un ou plusieurs phénomènes, description comprenant*

- *L'ensemble de tous les* événements *possibles. Cet ensemble est noté* $\mathfrak{S}$ *et les événements sont notés par des lettres majuscules* A*,* B*, ...*
- *Une* mesure de probabilité*, notée* p*.*
- *Un* univers*, noté* Ω*, précise comment on a choisi de « découper » le réel en possibles. Il doit y avoir une certaine cohérence dans la description des événements et de leurs probabilités. Par exemple certains événements ne peuvent coexister ; d'autres, au contraire, se produisent toujours ensemble, etc. C'est* Ω *qui décrit toutes les relations entre événements possibles.*

2 Un exercice

On lance trois dés. Donner un espace de probabilité $(\Omega, \mathfrak{S}, p)$ associé à cette expérience dans les trois cas suivants :

1) on distingue les trois dés,

2) on ne distingue pas les trois dés,

3) on ne s'intéresse qu'à la somme des trois dés.

1) Si l'on distingue les trois dés (par exemple on en lance un, puis le deuxième, puis le troisième ; ou bien on les lance tous les trois en même temps, mais on les distingue par leur couleur, etc.), le phénomène consiste en trois variables aléatoires : la valeur X_1 prise par le premier dé à l'issue du lancer, la valeur X_2 prise par le deuxième, et la valeur X_3 prise par le troisième. Chaque événement pourra être décrit en formulant certaines conditions portant sur ces variables X_1, X_2, X_3.

Considérons par exemple l'événement suivant :

$$A = \text{« on obtient au moins deux faces identiques »}$$

On pourra reformuler cet événement de la manière suivante :

$$A = \{X_1 = X_2 \quad \text{ou} \quad X_2 = X_3 \quad \text{ou} \quad X_3 = X_1\}$$

De même, l'événement $B =$ « on obtient trois faces identiques » peut aussi s'écrire :

$$B = \{X_1 = X_2 = X_3\}$$

Quand l'événement A se produit, il arrive parfois que B aussi se produise : quand deux faces sont identiques, il arrive parfois que l'autre leur soit aussi identique. D'ailleurs, à chaque fois que B se produit, *a fortiori* A se produit. Afin de mieux comprendre la relation entre ces deux événements, on va chercher à les décrire de manière explicite en énumérant toutes les valeurs des dés pour lesquelles ils se produisent. A l'issue d'un lancer, chacun des trois dés prend une valeur déterminée ; on peut noter ces trois valeurs par un triplet de nombres compris entre un et six (X_1, X_2, X_3). L'événement B se produit quand ces trois nombres sont égaux, c'est-à-dire pour les triplets suivants :

$$(1,1,1), (2,2,2), (3,3,3), (4,4,4), (5,5,5), (6,6,6)$$

Et l'événement A se produit pour les triplets suivants :

$(1,1,2),(1,1,3),...,(1,1,6),$ $(2,2,1),(2,2,3),...,(2,2,6),...$ $(6,6,1),(6,6,2),...,(6,6,5),$
$(1,2,1),(1,3,1),...,(1,6,1),$ $(2,1,2),(2,3,2),...,(2,6,2),...$ $(6,1,6),(6,2,6),...,(6,5,6),$
$(2,1,1),(3,1,1),...,(6,1,1),$ $(1,2,2),(3,2,2),...,(6,2,2),...$ $(1,6,6),(2,6,6),...,(5,6,6),$
$(1,1,1),$ $(2,2,2),...$ $(6,6,6).$

Les triplets décrivant B figurent tous parmi les triplets décrivant A : l'événement A se produit chaque fois que B se produit. Le fait de décrire les événements en énumérant tous les triplets de valeurs permet de constater les relations entre les événements. On a découpé le réel en événements possibles, et chaque triplet de valeurs décrit un *événement élémentaire*. Certes, le résultat $(1,1,2)$ ne dénote pas toutes les caractéristiques du réel : le deuxième dé peut avoir, ou ne pas avoir, roulé davantage que le troisième. On pourrait donc encore découper le réel en deux événements possibles : « les dés marquent $(1,1,2)$ et le deuxième a roulé davantage que le troisième », et « les dés marquent $(1,1,2)$ et le deuxième a moins roulé que le troisième ». Mais tout modèle est imparfait, et toute description du réel doit s'en tenir à certaines limites. Dans le cadre d'un pari, seule la valeur des dés nous intéresse, donc on ne cherchera pas à « découper » davantage le réel. L'univers Ω consiste en l'ensemble des triplets de nombres de 1 à 6. Chacun de ces triplets constitue un événement élémentaire. Décrire les événements en énumérant les événements élémentaires qui les composent permet de comprendre leurs relations.

Intéressons-nous maintenant à l'événement C décrit par les triplets suivants :

$$(6,6,2),(6,6,3),(6,6,4).$$

Combien C est-il probable ? Autrement dit : quelle valeur donnera-t-on à $p(C)$? Raisonnons de manière intuitive. C a d'autant plus de chances de se produire que les événements élémentaires le décrivant sont probables. Mais dans ce modèle, les événements élémentaires en lesquels on a choisi de « découper le réel » semblent avoir tous la même probabilité. En effet, on ne croit pas qu'il soit plus probable d'obtenir $(6,6,2)$ que $(1,2,3)$, ou que $(3,2,1)$, ou que $(1,1,1)$. A moins qu'un dé ne soit truqué ; mais en l'absence d'information à ce sujet, on doit concevoir le modèle comme si les dés étaient parfaits. Si l'univers Ω compte N événements élémentaires, on dira que « C a trois chances sur N de se produire ». Or $N = 6 \times 6 \times 6 = 216$. Donc :

$$p(C) = \frac{3}{216}$$

On sera donc souvent amené, pour déterminer la probabilité d'un événement, à compter des triplets. Pour ce faire, on s'aide parfois d'*arbres*. On dessine un arbre représentant les choix successifs à faire pour spécifier un des triplets parmi tous les triplets que l'on compte. Calculons par exemple $p(B)$. Combien y a-t-il de triplets dans l'événement B ? Pour spécifier un des triplets de B, on fait les choix représentés sur l'arbre (fig. 1 p. 5). L'arbre comprend une *racine* (à gauche) et des *feuilles* (extrémités des branches, à droite). Chaque feuille correspond à un triplet de B, et il suffit donc de compter les feuilles :

$$3 \times 6 \times 5 + 6 = 96$$

Donc $p(B) = \dfrac{96}{216}$.

2) Quel modèle choisir si on ne distingue pas les trois dés (si on les lance tous les trois en même temps et qu'ils sont indiscernables, de même couleur, ou bien qu'on ne souhaite pas les distinguer même s'ils sont de couleurs différentes en fait) ? On ne pourra plus distinguer non

plus les événements que l'on notait précédemment $(1,1,2)$, $(1,2,1)$, et $(2,1,1)$: il constitueront à eux trois un événement *élémentaire* « la valeur 2 apparaît deux fois et la valeur 1 une fois ». L'univers Ω contiendra trois types d'événements élémentaires :

(i) « on obtient trois valeurs distinctes p, q, r »

(ii) « la valeur p apparaît deux fois, et une autre valeur, q, apparaît une fois »

(iii) « les trois dés ont la même valeur p »

Remarquons qu'à présent, les événements élémentaires ne sont pas tous équiprobables. Par exemple, l'événement élémentaire

$$D = \text{« on obtient les trois valeurs 1, 2 et 3 »}$$

est plus probable que l'événement élémentaire

$$E = \text{« on obtient trois fois le chiffre 1 »}.$$

Dans l'ancien modèle, on aurait décomposé D en les six événements élémentaires suivants :

$$(1,2,3), (3,1,2), (2,3,1), (3,2,1), (1,3,2), (2,1,3)$$

On aurait donc $p(D) = \dfrac{6}{216} = \dfrac{1}{36}$. Tandis que l'événement E a seulement une chance sur 216 de se produire. Remarquez qu'on a naturellement recours à l'ancien modèle, *équiprobable*, pour déterminer les probabilités des événements élémentaires du nouveau modèle. C'est légitime : si la mesure de probabilité est un tant soit peu objective, il faut bien que les événements aient chacun la même probabilité dans les deux modèles.

3) Enfin, si on ne s'intéresse qu'à la somme des trois dés, l'univers des possibles est simplement l'ensemble des valeurs possibles de cette somme :

$$\{3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18\}$$

Ici non plus, le modèle n'est pas équiprobable : on sent bien qu'il est plus rare d'obtenir 18 que d'obtenir 10.

3 Des relations entre les événements

Dans la section précédente, on a vu sur un exemple comment découper le réel en événements élémentaires. On note Ω l'ensemble des événements élémentaires. On identifie alors chaque événement à un sous-ensemble de Ω, et les relations entre ces sous-ensembles décrivent bien les relations entre les événements correspondants ainsi qu'entre leurs probabilités respectives. Notons $\mathcal{P}(\Omega)$ l'*ensemble des parties de* Ω, c'est-à-dire l'ensemble de ses sous-ensembles. On a donc $\mathfrak{S} \subset \mathcal{P}(\Omega)$.

Conséquence immédiate de cette modélisation : deux événements sont égaux dès lors qu'ils sont composés des mêmes événements élémentaires.

Si $A \subset B$, alors l'événement A entraîne que l'événement B se réalise nécessairement, et $p(A) \leq p(B)$. L'événement Ω est certain, c'est-à-dire que $p(\Omega) = 1$, et l'événement $\emptyset$ est impossible, c'est-à-dire que $p(\emptyset) = 0$.

La conjonction de A et B, c'est-à-dire l'événement « A et B », correspond à l'intersection $A \cap B$.

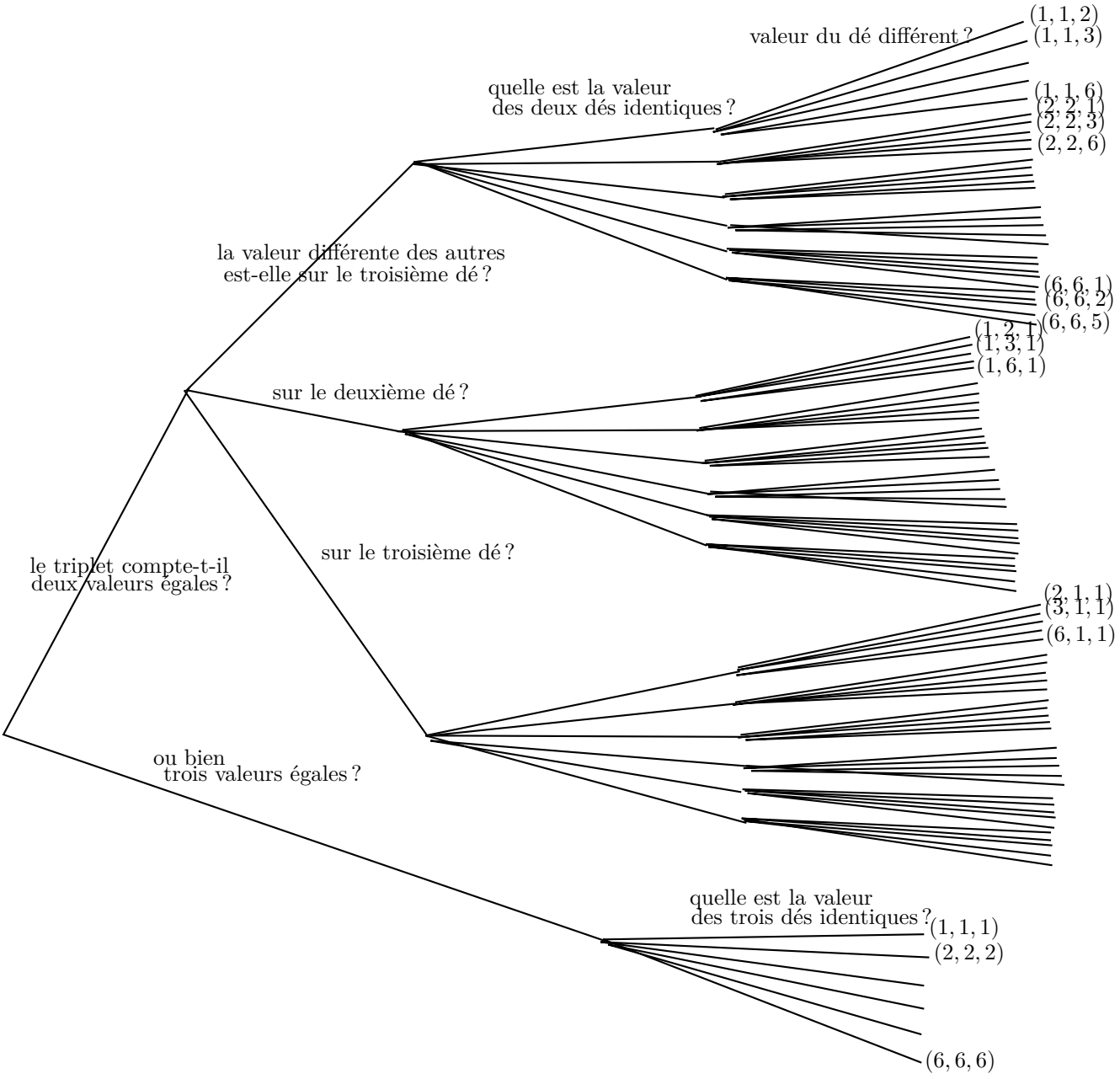


FIGURE 1 – Arbre combinatoire

La disjonction de deux événements A et B, c'est-à-dire l'événement « A ou B », correspond à la réunion $A \cup B$.

Si A et B sont des sous-ensembles *disjoints* de Ω (c'est-à-dire que $A \cap B = \emptyset$), alors les événements correspondants ne peuvent jamais se réaliser conjointement. On dit que ces événements sont *incompatibles*.

Théorème 3.1 *Si les événements A_1, A_2,... A_n sont deux-à-deux incompatibles, alors on a*[1]

$$p(A_1 \cup A_2 \cup ... \cup A_n) = p(A_1) + p(A_2) + ... + p(A_n)$$

Remarque. Soit des sous-ensembles $A_1, A_2, ..., A_n \subset \Omega$ deux-à-deux disjoints, tels que

$$\Omega = A_1 \cup A_2 \cup ... \cup A_n$$

On dit que $\{A_1, A_2, ..., A_n\}$ est une *partition* de Ω. Tout événement B peut alors s'écrire comme réunion disjointe de la manière suivante :

$$B = (B \cap A_1) \cup (B \cap A_2) \cup ... \cup (B \cap A_n)$$

On peut alors écrire (*formule de probabilité totale*) :

$$p(B) = p(B \cap A_1) + p(B \cap A_2) + ... + p(B \cap A_n)$$

Pour tout sous-ensemble $A \subset \Omega$, il existe un *complémentaire* $\complement_\Omega A$ correspondant à l'événement « non A », de sorte que $\{A, \complement_\Omega A\}$ est une partition de Ω. On a alors :

$$p(\complement_\Omega A) = 1 - p(A)$$

Exercice 3.1 [f1] *On lance trois dés. Calculer la probabilité d'avoir :*

1) au moins un six,

2) exactement un six,

3) au moins deux faces identiques,

4) au moins deux faces identiques et la somme des points paire,

5) au moins deux faces identiques ou la somme des points paire.

Exercice 3.2 [f1] *On truque un dé de telle sorte que la probabilité d'obtenir un numéro soit proportionnelle à ce numéro. Donner la probabilité de chaque numéro, celle d'obtenir un numéro pair, et celle d'obtenir un numéro premier.*

Exercice 3.3 [f1] *On considère un alphabet de n lettres.*

1) Calculer le nombre de mots de k lettres distinctes que l'on peut écrire avec cet alphabet.

2) Calculer la probabilité de tirer, parmi tous les mots de k lettres, un mot écrit avec des lettres toutes distinctes.

Exercice 3.4 [f1] *Quelle est la probabilité que parmi n personnes, deux personnes aient leur anniversaire le même jour.*

Exercice 3.5 [f1, III.176]

1. Ce théorème marche aussi pour des réunions infinies.

1) Soit $A_1, ..., A_n$ des ensembles finis, montrer que :

$$\mathrm{Card}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{Card} A_i - \sum_{1\leq i<j\leq n} \mathrm{Card}(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} \mathrm{Card}(A_i \cap A_j \cap A_k) + ... + (-1)^{n-1} \mathrm{Card}\left(\bigcap_{i=1}^{n} A_i\right)$$

2) Calculer le nombre de permutations de $\{1, ..., n\}$ qui ne fixent aucun point.

3) Un facteur répartit les quittances de loyer de n locataires au hasard, une dans chaque boîte. Calculer la probabilité p_n qu'aucun locataire n'ait reçu sa quittance. Calculer $\lim p_n$ (indication : appliquer la formule de Taylor en zéro à la fonction $t \mapsto e^{-t}$).

4) Calculer la probabilité que k locataires exactement aient reçu leurs quittances.

5) Un facteur distribue au hasard dans n boîtes p prospectus en oubliant au fur et à mesure où il en a placé. Calculer la probabilité que i boîtes soient vides.

4 Probabilités conditionnelles

On note $p_B(A)$ la probabilité que l'événement A se produise *sachant que* l'événement B est arrivé. Cela revient à « diviser tout » par $p(B)$, de sorte que la probabilité de l'événement B devienne 1 (puisqu'alors l'événement B est certain), c'est-à-dire à exclure de l'univers des possibles les situations dans lesquelles l'événement B n'est pas réalisé. Seules les situations où B se réalise sont à présent conçues comme des situations possibles : on ne considère plus que les possibles qui arrivent en même temps que B. Autrement dit :

$$p_B(A) = \frac{p(A \text{ et } B)}{p(B)}$$

On dit que deux événements A et B de probabilités non nulles sont *indépendants* si $p_A(B) = p(B)$. Cette définition se traduit aussi par l'égalité symétrique suivante :

$$p(A \text{ et } B) = p(A)p(B)$$

Remarque. Intuitivement, lorsque l'occurence de l'événement B n'a pas d'influence sur le caractère plus ou moins probable de l'événement A (aucun lien de causalité entre les deux événements), on peut dire que ces deux événements sont indépendants. C'est d'ailleurs ainsi qu'on procède, sans y penser, pour déterminer la probabilité d'obtenir 2 puis 3 en lançant successivement deux dés. La probabilité d'obtenir 2 au premier lancer est $p(A) = 1/6$, la probabilité d'obtenir 3 au second *sachant que* l'on a obtenu 2 au premier est encore $p_A(B) = 1/6$, *car ces deux événements sont indépendants*. La probabilité de la conjonction des deux événements est alors $p(A \text{ et } B) = 1/36$.

Mais attention, si l'absence de lien causal entraîne l'indépendance de deux événements, le contraire n'est pas toujours vrai. Quand la description des deux événements ne permet pas d'affirmer de manière évidente qu'il n'y aucun lien entre eux, il faut toujours vérifier l'indépendance en calculant les probablités de A, de B et de « A et B ».

Si les événements A, B, C,... constituent une partition de l'univers des possibles, on a :

$$p(E) = p(E \mid A)p(A) + p(E \mid B)p(B) + p(E \mid C)p(C) + ...$$

Soit deux événements A et B. Ils déterminent une partition de l'ensemble des possibles en quatre événements disjoints : « A et B », « A et non B », « non A et B » et « non A et non B ». Il est commode de représenter cette partition par un arbre, en indiquant sur chaque branche les

probabilités conditionnelles (fig. 2 p. 11). Attention à ne pas confondre ces arbres avec les arbres combinatoires utilisés plus haut pour dénombrer les éléments d'un ensemble.

Attention à ne pas confondre les événements *incompatibles*, pour lesquels $p_B(A) = 0$, et les événements *indépendants*, pour lesquels $p_B(A) = p(A)$.

Formule de Bayes.

Exercice 4.1 [IV.20] *Un lycée compte chaque année 25 inscrits dans une unique classe de terminale. On regarde les statistiques suivantes sur les reçus au baccalauréat en 2007 :*

	inscrits	*reçus*
non redoublants	22	12
redoublants	3	3

On dispose aussi des statistiques pour 2008 :

	inscrits	*reçus*
non redoublants	15	8
redoublants	10	9

Après le bac en 2007, si l'on choisit au hasard l'un des 25 élèves l'ayant passé, quelle est la probabilité qu'il ait eu le bac ? La probabilité qu'il ait eu le bac sachant qu'il a redoublé ? La probabilité qu'il ait eu le bac sachant qu'il n'est pas redoublant ? Même question pour 2008. Comparer les résultats avec ceux de 2007.

Exercice 4.2 [f1] *Soit $(\Omega, \mathcal{A}, p)$ un espace de probabilité. On considère deux événements A et B, et $\{C_i\}$ une partition de Ω.*

1) Calculer $p(B|A)$ en fonction de $p(A)$, $p(B)$ et $p(A|B)$.

2) Connaissant $p(B)$, $p(A|B)$, $p(A|C_i)$ et $p(C_i)$, calculer $p(B|A)$.

Application : soit p_i la probabilité pour qu'un couple ait exactement i enfants, calculer la probabilité pour qu'un couple ait un enfant unique sachant qu'il n'a pas de fille.

Exercice 4.3 [f1] *Pour les familles ayant trois enfants, les deux événements suivants sont-ils indépendants ?*

A = la famille a, à la fois, au moins une fille et au moins un garçon

B = la famille a au plus une fille

Et pour les familles à quatre enfants ?

Exercice 4.4 [f1] *Mon voisin a deux enfants dont au moins une fille, quelle est la probabilité que l'autre soit un garçon ? Un autre voisin a deux enfants dont le plus jeune est une fille, quelle est la probabilité que l'autre soit un garçon ?*

Exercice 4.5 [f1] *On a décelé dans un élevage de moutons une probabilité de $0,3$ pour qu'un animal soit atteint par une maladie M (en biostatistiques, on appelle ça la* prévalence *de la maladie).*

1. *(a) On choisit au hasard un animal de l'élevage. Quelle est la probabilité qu'il soit malade ?*
   (b) On choisit successivement et au hasard dix animaux. On appelle X la variable aléatoire égale au nombre d'animaux malades parmi eux. Montrer que X suit une loi binomiale dont on donnera les paramètres. Calculer son espérance mathématique.

(c) *Calculer* $p($« *aucun animal n'est malade parmi les dix* »$)$. *Puis calculer* $p($« *au moins un animal est malad*

2. *La probabilité qu'un mouton qui n'est pas atteint par* M *ait une réaction négative à un test* T *est* $0,9$ *(c'est la* spécificité *du test). S'il est atteint par* M*, la probabilité qu'il ait une réaction positive à* T *est* $0,8$ *(c'est la* sensibilité *du test).*
   (a) *Représenter par un arbre pondéré les données de l'énoncé.*
   (b) *Calculer* $p(T)$.
   (c) *Quelle est la probabilité qu'un mouton pris au hasard et ayant une réaction positive soit atteint par* M *?*

Exercice 4.6 [f1] *En phase finale d'un jeu télévisé, il y a une voiture à gagner. Le présentateur montre trois portes, derrière une porte il y a la voiture et derrière les deux autres un cochon. Il demande au candidat de choisir une porte. Le présentateur ouvre alors l'une des deux autres portes ; il y a un cochon derrière (et tout le monde le voit). Il reste donc deux portes derrière lesquelles la voiture puisse être. Le candidat a-t-il intérêt à conserver son choix ou bien devrait-il au contraire changer d'avis et ouvrir l'autre porte ?*

Exercice 4.7 [f1] *Un blog sur Internet rapporte l'existence d'un « test rapide » pour dépister la grippe H1N1. Le blog annonce que la sensibilité du test est 63 %, et sa spécificité 100 %. Une autre étude fournit une estimation de la « prévalence » du virus : au sein d'une population d'individus présentant des symptômes respiratoires, la proportion d'individus vraiment atteints par le virus est de 4 %.*

On rappelle que la spécificité *du test est la probabilité qu'un individu qui n'est pas atteint par le virus ait une réaction négative au test. La* sensibilité *du test est la probabilité qu'un individu qui est atteint par le virus ait une réaction positive au test.*

1. (a) *On choisit au hasard un individu présentant des symptômes respiratoires. Quelle est la probabilité qu'il ait la grippe H1N1 ?*
   (b) *On choisit successivement et au hasard dix individus présentant des symptômes respiratoires. On appelle* X *la variable aléatoire égale au nombre d'individus atteints par le virus parmi eux. Montrer que* X *suit une loi binomiale. Calculer l'espérance de* X.
   (c) *Calculer* $p($« *aucun individu n'est malade parmi les dix* »$)$. *Puis calculer* $p($« *au moins un individu est malade parmi les dix* »$)$.
2. (a) *On choisit au hasard un individu présentant des symptômes respiratoires ; on le soumet au « test rapide ». Quelle est la probabilité qu'il réagisse positivement ?*
   (b) *Quelle est la probabilité qu'un individu présentant des symptômes respiratoires* et *réagissant positivement au test soit atteint par le virus ?*
3. *Un lecteur du blog commente :*

Si la spécificité est vraiment de 100 %, c'est parfait pour l'épidémiologie. Seuls les résultats positifs comptent.
Mais si je suis médecin :
- *Soit il y a peu de grippes H1N1 comme maintenant (4 % des infections respiratoires aiguës) : la probabilité que mon malade ait la grippe est déjà de 4 % seulement ! Il faut un test ? Je ne vois pas trop l'utilité.*
- *Soit la grippe H1N1 devient très fréquente et là :*
  - *test positif, je traite comme une grippe pandémique*
  - *il est négatif, je ne peux pas le croire, donc je traite et isole comme une grippe pandémique.*
  - *j'explique au malade que le test négatif ne permet pas d'écarter le diagnostic : il peut me demander « alors pourquoi me l'avoir fait ? à quoi a servi le test ? ». On peut se faire accuser de facturer des tests inutiles...*

Expliquez la pensée du lecteur en vous aidant des questions précédentes.

Exercice 4.8 [III.152-155, f1] *Un livre contient quatre erreurs. A chaque relecture, chaque faute non corrigée est corrigée avec probabilité* $1/3$. *Les relectures sont indépendantes les unes des autres.*

1. *Combien faut-il faire de relectures pour que la probabilité qu'il ne subsiste aucune erreur soit supérieure à* $0,9$ *?*
2. *Traiter la même question en supposant que le nombre* x *d'erreurs est réparti de manière équiprobable sur* $\{0,1,2,3,4\}$.

Indication : *il faut s'intéresser à l'événement* $A_k =$ *« l'erreur A a été corrigée en au plus* k *lectures ».*

Exercice 4.9 [f1] *On choisit au hasard un des nombres entiers* $1,2,...,n$, *tous les choix étant équiprobables. Soit* $p \leq n$ *un entier non nul, et* A_p *l'événement « le nombre choisi est divisible par* p *».*

1) Calculer $p(A_p)$ *lorsque* p *divise* n.

2) Montrer que si $p_1,...,p_k$ *sont des diviseurs premiers distincts de* n, *alors les événements* $A_{p_1},...,A_{p_k}$ *sont indépendants.*

3) On appelle fonction indicatrice d'Euler la fonction Φ *définie sur les entiers naturels et dont la valeur* $\Phi(n)$ *est égale au nombre d'entiers inférieurs à* n *et premiers avec* n. *Montrer que :*

$$\frac{\Phi(n)}{n} = \prod_{p \text{ premier, } p|n} \left(1 - \frac{1}{p}\right)$$

5 Variables aléatoires discrètes

Soit E un ensemble et X une fonction de la forme

$$\begin{array}{rl} X : & \Omega \longrightarrow E \\ & \omega \longmapsto X(\omega) \end{array}$$

On dit que X est une *variable aléatoire*. Il est alors commode de concevoir X comme une variable (en oubliant la dépendance fonctionnelle par rapport à Ω) pour décrire symboliquement les

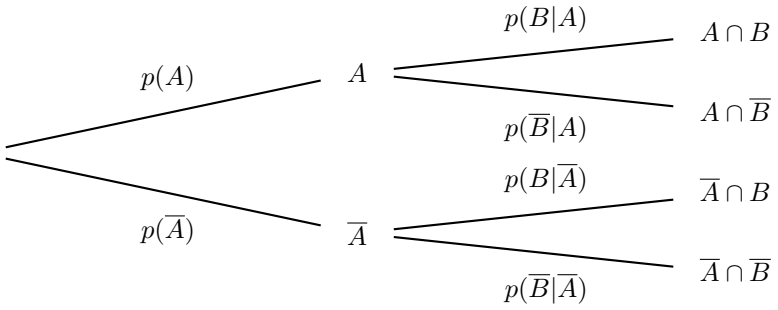


FIGURE 2 – Arbre conditionnel

événements. Par exemple, dans le premier modèle du lancer de trois dés où l'on a défini trois variables aléatoires X_1, X_2, X_3, on a vu qu'on pouvait décrire les événements A et B au moyen de cette notation :

$$A = \{X_1 = X_2 \quad \text{ou} \quad X_2 = X_3 \quad \text{ou} \quad X_3 = X_1\}$$
$$B = \{X_1 = X_2 = X_3\}$$

Soit l'événement $D =$ « le premier dé est impair ». On pourrait le noter

$$D = \{X_1 \in \Gamma\},$$

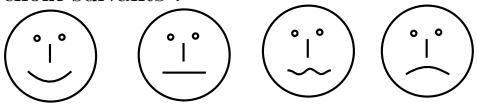
où $\Gamma = \{1, 3, 5\}$. En réalité, dans cet exemple, les X_k sont trois fonctions :

$$\begin{array}{rccc} X_k : & \Omega & \longrightarrow & \{1,2,3,4,5,6\} \\ & (i_1, i_2, i_3) & \longmapsto & i_k \end{array}$$

Alors D est l'image inverse $X_1^{-1}(\Gamma)$ de l'ensemble des nombres impairs Γ par la fonction X_1.

Si l'ensemble E est dénombrable (par exemple un ensemble de cardinal fini, ou bien l'ensemble des entiers relatifs $\mathbb{Z}$), on dit que X est une variable *discrète*. Au contraire, si $E = \mathbb{R}$, il s'agit d'une variable *continue*. Parmi les variables discrètes, si l'ensemble E est un ensemble de nombres, on dit que X est une variable *quantitative* ; sinon, il s'agit d'une variable *qualitative*. Parmi les variables discrètes qualitatives, on distingue encore les variables *catégorielles* des variables *ordonnées* pour lesquelles l'ensemble E est un ensemble ordonné.

Exemple 1 [IV.161] Lors d'une enquête de satisfaction des malades hospitalisés dans un certain hôpital, au cours d'une année donnée, chaque sujet doit cocher une case d'un questionnaire parmi les choix suivants :



Dans la pile des questionnaires archivés cette année-là, si l'on prend un questionnaire au hasard, la case cochée est une variable aléatoire discrète qualitative ordonnée. L'ensemble des valeurs de cette variable est en effet un ensemble ordonné.

Exemple 2 [IV.162] Chaque gène est codé dans la molécule d'ADN. Ainsi, dans le génome humain, l'exon 21 qui code une partie du gène MDAR a trois génotypes (c'est-à-dire trois valeurs possibles suivant les individus) CC, CT et TT. L'exon 26 qui en code une autre partie a cinq

génotypes GG, GA, AA, GU, AU. Dans la population caucasienne, la répartition des génotypes est la suivante :

	GG	GA	AA	GU	AU
CC	$0,60$	$0,08$	$0,02$	$0,00$	$0,00$
CT	$0,08$	$0,08$	$0,02$	$0,01$	$0,01$
TT	$0,02$	$0,04$	$0,02$	$0,01$	$0,01$

Si l'on prend un individu au hasard au sein de la population caucasienne, le génotype de l'exon 26 est une variable aléatoire *discrète qualitative catégorielle*, notons-la E_{26}. De même, notons E_{21} le généotype de l'exon 21. On peut alors écrire par exemple :

$$\begin{aligned} & p([E_{26} = AA] \ \cap \ [E_{21} = CC]) \\ & = p([E_{26} = AA] \ \cap \ [E_{21} = CT]) \\ & = p([E_{26} = AA] \ \cap \ [E_{21} = TT]) \\ & = 0,02 \end{aligned}$$

Donc

$$p(E_{26} = AA) = 0,06$$

Remarque 5.1 *On pourrait être tenté de concevoir une variable qualitative ordonnée comme variable quantitative en notant par un nombre chacune de ses valeurs, dans l'ordre. Ainsi dans le premier exemple, la satisfaction des malades pourrait être notée de 1 à 4. Pourtant, il semble peu raisonnable d'adopter une telle notation car la satisfaction, état émotionnel, n'est pas mesurable de manière intrinsèque par une quantité. Si l'on choisissait pourtant de le faire, on pourrait calculer la « satisfaction moyenne » en faisant la moyenne de ces notes : mais justement une telle moyenne n'a pas beaucoup de sens.*

Exemple 3 On lance un dé plusieurs fois, jusqu'à obtenir 6. On note X le nombre de lancers qu'il a fallu faire avant le premier 6. Attention, X semble être une variable quantitative, mais quelle est donc sa valeur si l'on lance le dé éternellement sans jamais obtenir de 6 ? En théorie, il est bien possible qu'un tel événement se réalise parfois. Il faut donc plutôt concevoir X comme une variable à valeur dans l'ensemble

$$\mathbb{N} \cup \{+\infty\}$$

Ce n'est donc pas, à proprement parler, une variable quantitative.

Définition 5.1 *On dit que deux variables aléatoires discrètes X et Y sont* indépendantes *si :*

$$(\forall (a,b) \in X(\Omega) \times Y(\Omega)) \quad p(X = a \ \wedge Y = b) = p(X = a)p(Y = b)$$

.

Définition 5.2 *Soit X et Y deux variables aléatoires quantitatives discrètes. On définit alors l'espérance* $\mathrm{E}(X)$*, la variance* $\mathrm{V}(X)$*, l'écart-type* σ_X *et la covariance* $\mathrm{cov}(X,Y)$ *:*

$$\mathrm{E}(X) = \sum_{a \in X(\Omega)} a \ p(X = a)$$

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right)$$

$$\sigma_X = \sqrt{V(X)}$$

$$\mathrm{cov}(X,Y) = \mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right)$$

Propriété 5.1 *Soit X et Y deux variables aléatoires quantitatives discrètes. L'espérance vérifie les propriétés suivantes.*

(i) L'espérance est linéaire. Soit a et b des constantes, on a

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

(ii) L'espérance est positive, c'est-à-dire que

$$[p(X \geq 0) = 1] \Rightarrow [\mathrm{E}(X) \geq 0]$$

(iii) Soit f une fonction réelle convexe. On note $f(X)$ la fonction composée :

$$\begin{array}{rccc} f(X): & \Omega & \longrightarrow & f \circ X(\Omega) \\ & \omega & \longmapsto & f \circ X(\omega) \end{array}$$

Alors

$$f(\mathrm{E}(X)) \geq \mathrm{E}(f(X))$$

(iv) Si X et Y sont indépendantes, on a

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$$

Propriété 5.2 *Si X et Y sont deux variables aléatoires quantitatives* indépendantes*, et a et b des constantes, on a :*

$$\mathrm{V}(aX + bY) = a^2V(X) + b^2V(Y)$$

En particulier, sous cette hypothèse, $\mathrm{V}(X + Y) = \mathrm{V}(X - Y) = \mathrm{V}(X) + \mathrm{V}(Y)$.

Définition 5.3 *Soit X une variable aléatoire discrète à valeurs dans un ensemble E. On définit la fonction suivante, appelée* loi de probabilité *de X :*

$$\begin{array}{rccc} \pi_X: & E & \longrightarrow & [0,1] \\ & a & \longmapsto & p(X = a) \end{array}$$

Deux variables aléatoires admettant la même loi sont dites « équidistribuées ».

Définition 5.4 *Soit X une variable aléatoire quantitative. On définit la fonction suivante, appelée* fonction de répartition *de X :*

$$\begin{array}{rccc} F: & \mathbb{R} & \longrightarrow & [0,1] \\ & a & \longmapsto & p(X \leq a) \end{array}$$

Exercice 5.1 [f1] *Décrire l'espace de probabilité de l'expérience aléatoire qui consiste à répartir au hasard r boules dans n cases (pour chacune des boules, on choisit au hasard, avec la même probabilité, l'une quelconque des n cases). Calculer la loi de probabilité notée $\mu_{r,n}$ du nombre de boules tombant dans une case donnée à l'avance et montrer que, si r et n tendent vers $+\infty$ de sorte que r/n tend vers un réel $\lambda > 0$, alors $\mu_{r,n}(k)$ tend vers :*

$$\mu(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

Exercice 5.2 [IV.15-17] *Un jeu entre deux partenaires est dit équitable si les espérances de gain de chaque joueur, c'est-à-dire les moyennes de leurs gains considérés comme variables aléatoires, sont égales. Dans ce qui suit les gains sont constitués par les récompenses (supposées égales) du joueur gagnant.*

a) Jeu de « passe sept ». On lance deux dés. Pierre gagne si le total des points dépasse 7*, et perd s'il n'atteint pas* 7 *(s'il égale* 7 *la partie est nulle). Ce jeu est-il équitable ?*

b) Jeu de « passe dix ». On lance trois dés. Pierre gagne si le total des points dépasse 10 *et perd s'il ne dépasse pas* 10*. Le jeu est-il équitable ?*

c) Calculer dans chacun des deux cas précédents la loi de la variable aléatoire « total des points », son espérance et sa variance.

Exercice 5.3 [I.74] *Soit X une variable aléatoire discrète telle que $X(\Omega) = \mathbb{N}^\star$. Montrer que la loi suivante est bien une loi de probabilité :*

$$P(X = n) = \frac{4}{n(n+1)(n+2)}$$

Exercice 5.4 [I.80-81] *Déterminer a pour que les égalités suivantes définissent une loi de probabilité :*

$$\forall n \in \mathbb{N}, p(X = n) = \frac{a(n+3)}{2^n}$$

Calculer $\mathrm{E}(X)$, $\mathrm{E}(X^2)$ et $\mathrm{V}(X)$.

Exercice 5.5 [I.81] *Soit $x \in]0,1[$. Déterminer a pour que la loi suivante soit une loi de probabilité :*

$$\forall n \in \mathbb{N}, P(X = n) = \frac{x^n}{n+1}$$

Déterminer son espérance et sa variance.

Exercice 5.6 [I.81-82] *On lance un dé trois fois de suite : on obtient des valeurs a, b et c. On considère l'équation $ax^2 + bx + c$. On définit une variable aléatoire X par $X = 1$ si cette équation admet deux racines réelles distinctes, $X = 0$ si elle admet une racine double, et $X = -1$ si elle n'admet pas de racine réelle. Déterminer la loi de probabilité de X. Calculer son espérance et sa variance.*

Exercice 5.7 [I.83] *Trouver toutes les lois de probabilité $(p_n)_{n\in\mathbb{N}}$ formant une suite arithmético-géométrique. Déterminer à chacune son espérance et sa variance.*

Exercice 5.8 [I.84-86] *Trouver toutes les lois de probabilité $(p_n)_{n\in\mathbb{N}}$ telles que pour tout $n \in \mathbb{N}$,*

$$p_{n+2} = \frac{5p_{n+1} - p_n}{6}$$

On pourra expliciter p_n en fonction de n et de p_0 et p_1, exprimer p_1 en fonction de p_0, et montrer que $p_0 \leq 1/3$; réciproquement, toute valeur de p_0 telle que $p_0 \leq 1/3$ convient.

Exercice 5.9 [I.87] *On remplace une carte quelconque (autre que l'as de pique) d'un jeu de trente-deux cartes par un deuxième as de pique. Soit n et p deux entiers strictement positifs.*

1. *On tire sans remise n cartes. Probabilité de déceler la supercherie.*
2. *Ici, $n = 4$. On recommence p fois le tirage de la question précédente, en remettant à chaque fois les quatre cartes tirées dans le jeu entre chaque tirage. Quel est le nombre d'expériences nécessaire pour déceler la supercherie avec une probabilité supérieure à $0,95$?*

Exercice 5.10 *Soit E un ensemble et F un sous-ensemble de E. L'indicatrice de F est la fonction $\mathbb{1}_F$ définie par :*

$$\begin{array}{rccl} \mathbb{1}_F : & E & \longrightarrow & \{0,1\} \\ & x & \longmapsto & \mathbb{1}_F(x) = \left\{ \begin{array}{l} 1 \text{ si } x \in F \\ 0 \text{ sinon} \end{array} \right. \end{array}$$

a) Démontrez les identités suivantes :

$$\mathbb{1}_{A\cap B} = \mathbb{1}_A.\mathbb{1}_B$$

$$\mathbb{1}_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A.\mathbb{1}_B$$

b) Soit X une variable aléatoire à valeurs dans E et Γ un sous-ensemble de E. Démontrez que

$$\mathbb{1}_{X\in\Gamma} = \mathbb{1}_\Gamma \circ X$$

c) Soit X une variable aléatoire à valeurs numériques avec $X(\Omega) = \{a_1, a_2, ..., a_n\}$ où les a_i sont des réels distincts. Démontrez que X peut alors s'écrire comme combinaison linéaire d'indicatrices :

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{X=a_i}$$

6 Un peu d'analyse combinatoire

Permutations Une permutation de n objets $a_1, a_2, ..., a_n$ est un arrangement de ces n objets dans un certain ordre. Les n objets doivent être distincts. Par exemple, voici deux permutations :

$$a_1\ a_2\ a_3\ ...\ a_n$$

$$a_2\ a_1\ a_3\ ...\ a_n$$

On compte $n!$ permutations de n objets.

La situation est différente si certains des n objets sont inderscernables. Par exemple, si $a_1 = a_2$, les deux permutations ci-dessus sont en fait égales. Plus généralement, si parmi n objets on compte p groupes distincts d'objets tous semblables, le premier groupe comptant α_1 objets, le second α_2, etc., alors le nombre de permutations sera

$$\frac{n!}{\alpha_1!\alpha_2!...\alpha_p!}$$

Par exemple, le nombre de mots formés avec les six lettres E, M, M, E, L, E, est $\dfrac{6!}{3!2!} = 60$.

Arrangements Un arrangement de p objets parmi n objets (distincts) est obtenu en choisissant p objets parmi ces n objets et en les rangeant dans un certain orbre. On en compte

$$A_n^p = \frac{n!}{(n-p)!} = n(n-1)(n-2)...(n-p+1)$$

Combinaisons Une combinaison de p objets parmi n objets (distincts) est obtenue en choisissant p objets parmi ces n objets. On en compte

$$C_n^p = \frac{n!}{p!(n-p)!}$$

Quelques formules utiles

$$C_n^0 = C_n^n = 1, \quad C_n^1 = n, \quad C_n^{n-p} = C_n^p$$

$$p!C_n^p = A_n^p, \quad kC_n^k = nC_{n-1}^{k-1}$$

$$C_{n+1}^{p+1} = C_n^{p+1} + C_n^p, \quad 2^n = \sum_{k=0}^{n} C_n^k$$

$$C_{n+1}^{p+1} = C_n^p + C_{n-1}^p + C_{n-2}^p + ... + C_p^p$$

Exercice 6.1 *On distribue cinquante-deux cartes à quatre personnes, treize pour chacune. Calculer le nombre de distributions possibles. Quelle est la probabilité pour un joueur donné d'avoir :*

a) un roi ?

b) au moins un roi ?

c) un as et un roi ?

d) au moins un as et au moins un roi ?

Exercice 6.2 *Une boite contient huit boules rouges, trois jaunes et neuf bleues. Si l'on tire au hasard et sans les replacer trois de ces boules, évaluez la probabilité de chacun des événements suivants :*

a) Les trois boules sont rouges.

b) Au moins une boule est jaune.

c) Deux boules sont rouges, une jaune.

d) Il y a une boule de chaque couleur.

e) Les boules tirées sont, dans l'ordre, rouge puis jaune puis bleue.

f) Il y a au moins une boule bleue et au moins une boule rouge.

7 Quelques lois discrètes

Loi de Bernoulli Si X ne prend que deux valeurs possibles a et b, $\text{Card}\,(X(\Omega)) = 2$, on dit que X suit une *loi de Bernoulli*. Remarquez qu'on a alors

$$p(X = a) = 1 - p(X = b)$$

Une telle loi permet par exemple de modéliser un unique tirage aléatoire à « pile ou face ».

Propriété 7.1 *Soit X une variable suivant une loi de Bernoulli de valeurs 0 et 1 :*

$$X = \begin{cases} 1 \text{ avec probabilité } \alpha \\ 0 \text{ avec probabilité } 1-\alpha \end{cases}$$

On a alors $\text{E}(X) = \alpha$ *et* $\text{V}(X) = \alpha(1-\alpha)$.

Propriété 7.2 *Si A est un événement, la fonction indicatrice $\mathbb{1}_A$ définie dans l'exercice 5.10 p. 15 est une variable aléatoire, et elle suit une loi de Bernoulli de paramètre $p(A)$.*

Loi uniforme sur un intervalle d'entiers $[\![a, b]\!]$ On dit que U suit une loi uniforme discrète sur l'intervalle $[\![a, b]\!]$ si

$$(\forall n \in [\![a, b]\!]) \quad p(U = n) = \frac{1}{b - a + 1}$$

On a alors $\mathrm{E}(U) = \dfrac{a+b}{2}$ et $\mathrm{V}(U) = \dfrac{(b-a)(b-a+2)}{12} = \dfrac{(b-a+1)^2 - 1}{12}$. Le « générateur de nombres aléatoires » d'un ordinateur simule une telle loi. Par exemple, la fonction `rand()` en langage C se comporte comme une variable aléatoire uniforme sur l'intervalle d'entiers $[\![0, \texttt{RAND\_MAX}]\!]$.

Loi binomiale Si on fait n expériences de type Bernoulli semblables et telles que la réalisation de chacune n'influe pas sur les autres, on aura des variables deux-à-deux indépendantes X_1, X_2,... X_n. Chacune de ces variables suit une même loi de Bernoulli à valeurs dans un ensemble à deux éléments $\{a, b\}$. Comptons le nombre de ces variables prenant la valeur a et notons-le Y. C'est une nouvelle variable aléatoire, que l'on peut décrire ainsi :

$$Y(\omega) = \mathrm{Card}\,\{i \in [1, n] \mid X_i(\omega) = a\}$$

Pour chaque variable X_k, on note $\alpha = p(X_k = a)$. On vérifiera aisément que, pour tout $0 \leq k \leq n$, on a :

$$p(Y = k) = C_n^k \alpha^k (1 - \alpha)^{n-k}$$

On dit que Y suit une *loi binomiale* de paramètres n, α.

Propriété 7.3 *Si Y suit une loi binomiale de paramètres n, α, on a :*

$$\mathrm{E}(Y) = n\alpha$$

$$\mathrm{V}(Y) = n\alpha(1 - \alpha)$$

Démonstration. Il suffit de remarquer qu'une telle variable peut s'écrire comme somme de variables indépendantes qui suivent toutes une même loi de Bernoulli :

$$Y = \sum_{i=1}^{n} \mathbf{1}_{[X_i = a]}$$

On utilise alors la linéarité de l'espérance... *Autre méthode :* le calcul direct de l'espérance et de la variance demande quelques connaissances algébriques sur les coefficients binomiaux. Par exemple, pour calculer l'espérance, on peut utiliser la propriété suivante :

$$kC_n^k = nC_{n-1}^{k-1}$$

On a alors :

$$\mathrm{E}(Y) = \sum kC_n^k \alpha^k (1-\alpha)^{n-k} = n\alpha \sum C_{n-1}^{k-1} \alpha^{k-1} (1-\alpha)^{(n-1)-(k-1)} = n\alpha(\alpha + 1 - \alpha)^{n-1} = n\alpha$$

Loi de Poisson On dit qu'une variable Z à valeurs entières positives suit une loi de Poisson de paramètre λ si

$$(\forall k \in \mathbb{N}) \quad p(Z = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Exemple : voir exercice 5.1 p. 13.

Propriété 7.4 *Si Z suit une loi de Poisson de paramètre λ, on a*

$$\mathrm{E}(Z) = \mathrm{V}(Z) = \lambda$$

Exercice 7.1 [IV.13-14] *On lance une fusée vers Saturne. On admet que la probabilité de succès est $0,7$. On décide de lancer des fusées jusqu'à ce que trois succès soient réalisés. Probabilité que cela nécessite moins de dix lancers ? Quelle devrait être la probabilité de succès pour que la probabilité d'avoir trois succès en moins de dix lancers soit supérieure à $0,95$?*

Exercice 7.2 *Une compagnie de transport désire optimiser les contrôles afin de limiter l'impact des fraudes et pertes occasionnées par cette pratique. Cette compagnie effectue une étude basée sur deux trajets par jour pendant les vingt jours ouvrables d'un mois soit au total quarante trajets. On admet que les contrôles sont indépendants les uns des autres et que la probabilité pour tout voyageur d'être contrôlé est égale à α. Le prix de chaque trajet est de dix euros, en cas de fraude l'amende est de cent euros.*

Claude fraude systématiquement lors des quarante trajets soumis à cette étude. Soit X_i la variable aléatoire qui prend la valeur 1 si Claude est contrôlé au i-ème trajet et la valeur 0 sinon. Soit X la variable aléatoire définie par $X = X_1 + X_2 + ... + X_{40}$.

1. *Déterminer la loi de probabilité de X, puis calculer $\mathrm{E}(X)$ et $p(X \leq 2)$.*
2. *Soit Z le gain algébrique (c'est-à-dire positif en cas de gain et négatif en cas de perte) réalisé par le fraudeur. Remarquer que Z est fonction affine de X et en déduire son espérance.*
3. *Remarquer que $p(X \leq 2)$ est une fonction de α, soit $f(\alpha)$. Montrer que f est strictement décroissante sur $[0,1]$. Résoudre par approximation l'équation*

$$f(\alpha) = 0,01$$

En déduire la valeur minimale qu'il faut attribuer à p afin que la probabilité que Claude subisse au moins trois contrôles soit supérieure ou égale à 99 %.

Exercice 7.3 [f2] *Soit $(\Omega, \mathcal{A}, p)$ un espace de probabilité, A et B deux événements de $\mathcal{A}$. On définit la variable aléatoire X par :*

$$X(\omega) = a\mathbf{1}_A(\omega) + b\mathbf{1}_B(\omega),$$

où a et b sont des réels non nuls.

1) Calculer la loi de X.

2) Que se passe-t-il si A et B sont incompatibles (ou disjoints) ? Et s'ils sont indépendants ?

3) Calculer l'espérance et la variance de X.

Exercice 7.4 [f2] *On joue à pile ou face avec une pièce non équilibrée. Soit X_i la variable aléatoire valant 1 si le i-ème lancer tombe sur pile et 0 sinon. On suppose que les lancers sont indépendants. Soit Y le nombre de faces obtenues avant le premier pile. Soit Y_1 et Y_2 deux variables aléatoires indépendantes et de même loi que Y.*

1) Montrer que, pour tout réel $u \neq 1$ et tout entier $n \geq 2$, on a les deux égalités suivantes :

$$\frac{1}{(1-u)^2} = 1 + 2u + 3u^2 + ... + nu^{n-1} + \frac{(n+1)u^n - nu^{n+1}}{(1-u)^2}$$

$$\frac{2}{(1-u)^3} = \sum_{k=2}^{n} k(k-1)u^{k-2} + \frac{n(n+1)u^{n-1} - 2(n-1)(n+1)u^n + n(n-1)u^{n+1}}{(1-u)^3}$$

2) Calculer la loi de Y, donner son espérance et sa variance.

3) Calculer $p(Y_1 = Y_2)$.

4) Calculer la loi de $Z = Y_1 + Y_2$.

5) Calculer la loi de $U = \inf(Y_1, Y_2)$.

Exercice 7.5 [Troe] *Soit $a, b \in \mathbb{N}^\star$ et X une variable aléatoire à valeurs dans $\{1, 2, ..., ab\}$ telle que pour tout $x \in \{1, 2, ..., ab\}$ on a ait*

$$P(X = x) = \frac{1}{a} - \frac{1}{b}$$

1. *Quelles conditions doivent vérifier a et b ?*
2. *Déterminer la fonction de répartition F de X et donner sa représentation graphique. Résoudre $F(u) = 1/2$.*
3. *Déterminer $\mathrm{E}(X)$, et trouver a et b tels que $\mathrm{E}(X) = 7/2$.*

Exercice 7.6 [f2] *On lance un dé à six faces. On note X le résultat. On lance alors un dé à X faces. Soit Y le résultat obtenu.*

1) Calculer la loi du couple (X, Y).

2) En déduire les lois de X et de Y.

3) Les variables X et Y sont-elles indépendantes ? Calculer $\mathrm{cov}(X, Y)$.

Exercice 7.7 [I.77] *k urnes numérotées de 1 à k contiennent chacune n boules numérotées de 1 à n. On extrait une boule de chaque urne. On note X_i la variable aléatoire égale au numéro de la boule tirée dans l'urne i. On définit la variable aléatoire Y par $Y = \max_{1 \leq i \leq k} X_i$.*

1. *Calculer la fonction de répartition de Y.*
2. *En déduire la loi de Y.*
3. *Calculer $\mathrm{E}(Y)$ et $\mathrm{V}(Y)$.*

Exercice 7.8 [I.78-79] *Soit X uen variable aléatoire suivant la loi de Pascal $\mathcal{P}(1, p)$ (c'est-à-dire que $\forall k \in \mathbb{N}, p(X = k) = p(1 - p)^k$). Soit $m \in \mathbb{N}$. On pose $Y = \min(X, m)$.*

1. *Quelles valeurs prend Y ?*
2. *Donner la loi de Y.*
3. *Calculer l'espérance et la variance de Y.*

Exercice 7.9 [f1] *Calculer la loi de $S_n = X_1 + ... + X_n$ où les X_i sont des variables aléatoires indépendantes et de loi de Poisson de paramètre λ_i. En déduire l'espérance et la variance de S_n.*

Exercice 7.10 [I.88] *Le nombre de visiteurs quotidiens d'un parc d'attraction suit une loi de Poisson de paramètre 10000. Ce parc a dix entrées E_1,...,E_{10} qui sont équiprobables.*

1. *Déterminer le nombre moyen de visiteurs en une journée.*
2. *On désigne par X_1 le nombre de visiteurs entrant par E_1 en une journée donnée. Déterminer la loi de X_1 et en déduire son espérance et sa variance.*
3. *Sachant qu'un visiteur sur dix se débrouille pour entrer sans payer, calculer le nombre moyen de visiteurs qui payent et entrent par E_1 par jour.*

Exercice 7.11 [I.89] *Une urne contient n jetons numérotés de 1 à n. On tire une poignée aléatoire éventuellement vide. On note Y le nombre de jetons, et X la somme des numéros obtenus. On suppose que Y suit une loi uniforme.*

1. *Préciser $X(\Omega)$.*
2. *Soit X_k la variable aléatoire égale à k si le jeton k est dans la poignée, et zéro sinon. Montrer que $\dfrac{X_k}{k}$ suit une loi de Bernoulli dont on précisera le paramètre.*
3. *Calculer $\mathrm{E}(X)$.*

Exercice 7.12 [I.90] *Soient α et β deux réels, et pour tout $k \in \mathbb{N}$*

$$p_k = a\frac{\alpha^k + \beta^k}{k!}$$

1. *Suivant les signes de α et β, discuter l'existence de a pour que $(p_k)_{k\in\mathbb{N}}$ soit une loi de probabilité. Le cas échéant, déterminer a.*
2. *Dans les cas où on a défini une loi, soit X une variable aléatoire réelle suivant cette loi. Déterminer l'espérance et la variance de X.*
3. *Peut-il arriver que X suive une loi de Poisson ?*

Exercice 7.13 [VII.61-67, To77] *Soit E un ensemble fini de cardinal n, et deux entiers $0 < p, q \leq n$. On choisit « au hasard » deux sous-ensemble F et G de E de cardinaux respectifs p et q.*

1. *Construire un modèle équiprobable de cette expérience : décrivez l'espace de probabilité Ω.*
2. *Que vaut $p(F \subset G)$? $p(F \cap G = \varnothing)$?*
3. *On pose $X = \mathrm{card}(F \cap G)$. Déterminer la loi de probabilité de la variable aléatoire X.*
4. *Soit n, p, q des entiers positifs avec $0 < p, q \leq n$, et Y une variable aléatoire dont la loi de probabilité est définie par :*

$$(\forall k \in [\![\max(p+q-n, 0), \min(p, q)]\!]) \quad p(Y = k) = \frac{C_p^k C_{n-p}^{q-k}}{C_n^q}$$

On dit alors que Y suit une loi hypergéométrique de paramètres n, p, q. Vérifiez qu'on a bien $\sum p(Y = k) = 1$; puis calculez $\mathrm{E}(Y)$.

5. *Montrez que*

$$\sum_{k=\max(p+q-n,0)}^{\min(p,q)} k^2 p(Y = k) = \frac{p(p-1)C_{n-2}^{q-2} + pC_{n-1}^{q-1}}{C_n^q}$$

Vérifiez alors que

$$\mathrm{V}(Y) = \frac{qp(n-q)}{n(n-1)}$$

6. Application numérique. *En utilisant le résultat de la question précédante, calculez $\mathrm{E}(X)$ dans le cas où $n = 6$, $p = 2$, $q = 3$.*
7. *On modifie l'expérience de la façon suivante : on fixe le sous-ensemble F une fois pour toutes, et on choisit au hasard un sous-ensemble G de cardinal q. Reprendre chacune des questions précédentes.*

8 Variable aléatoire fonction d'une autre variable aléatoire

Loi et espérance d'une fonction d'une variable aléatoire discrète Soit $X : \Omega \to \mathbb{E}$ une variable aléatoire et $f : E \to F$ une fonction. Alors la composée $f \circ X$, aussi notée $f(X)$, est une variable aléatoire. Si $X(\Omega)$ est un ensemble discret, X est une variable discrète et $f(X)$ aussi. Sa loi est alors décrite par

$$p(f(X) = y) = \sum_{x \in f^{-1}(y)} p(X = x)$$

Son espérance se calcule aisément :

$$\mathrm{E}(X) = \sum_y yP(f(X) = y) = \sum_x f(x)p(X = x)$$

Comment simuler une variable de Bernoulli Sur machine, on dispose souvent d'un générateur de nombres aléatoires qui simule une variable aléatoire discrète uniforme U à valeurs dans un intervalle d'entiers $[\![0, N]\!]$. Si N est divisible par q, il est alors facile de simuler une variable de Bernoulli de paramètre $\frac{p}{q}$. En effet, la composée de U et de la fonction $x \mapsto \begin{cases} 1 \text{ si } x < pN/q \\ 0 \text{ sinon} \end{cases}$ est une telle variable, aussi notée

$$\mathbb{1}_{U<pN/q}$$

Comment simuler une variable binomiale Dans les mêmes conditions, il est aussi possible de simuler une variable binomiale de paramètres $\left(n, \frac{p}{q}\right)$. Soit $U_1, U_2, ..., U_n$ des variables uniformes discrètes sur $[\![0, N]\!]$ deux à deux indépendantes, réalisées par n appels distincts du générateur de nombres aléatoires, alors la variable suivante est une variable binomiale de paramètres $\left(n, \frac{p}{q}\right)$:

$$\mathbb{1}_{U_1<pN/q} + \mathbb{1}_{U_2<pN/q} + ... + \mathbb{1}_{U_n<pN/q}$$

Loi d'un n-uplet et lois marginales Si $X_1, X_2, ..., X_n$ sont des variables aléatoires à valeurs dans $E_1, E_2, ..., E_n$ (respectivement), alors on note $(X_1, X_2, ..., X_n)$ la variable aléatoire définie par

$$\omega \longmapsto (X_1(\omega), X_2(\omega), ..., X_n(\omega)) \in E_1 \times E_2 \times ... \times E_n$$

La connaissance de la loi de cette variable n-uplet permet aisément de retrouver les lois des variables $X_1, X_2, ..., X_n$, dites *lois marginales*. En effet on a pour tout i :

$$p(X_i = x) = \sum_{k_1 \in E_1, ..., k_{i-1} \in E_{i-1}, k_{i+1} \in E_{i+1}, ..., k_n \in E_n} p\left((X_1, X_2, ..., X_n) = (k_1, ..., k_{i-1}, x, k_{i+1}, ..., k_n)\right)$$

En revanche, la connaissance des lois marginales ne suffit pas à déterminer la loi du n-uplet (sauf si les X_i sont deux à deux indépendants).

9 Loi des grands nombres

Soit X une variable aléatoire quantitative, et X_1, X_2,..., X_n des variables aléatoires indépendantes de même loi que X. On pose

$$Z = \frac{X_1 + X_2 + ... + X_n}{n}$$

Alors Z est une variable aléatoire. Mais pour n grand, Z est une bonne estimation de la l'espérance $\mathrm{E}(X)$. En effet :

$$\mathrm{E}(Z) = \mathrm{E}(X)$$
$$\sigma_Z = \frac{\sigma_X}{\sqrt{n}}$$

C'est-à-dire que pour n grand, $\sigma_Z \simeq 0$ et les valeurs prises par la variable Z sont très peu dispersées, elles se concentrent autour de son espérance $\mathrm{E}(X)$. Cette remarque donne un sens statistique au concept d'espérance. Le « théorème central limite » offrira une formulation plus précise de la loi des grands nombres : il nous montrera que pour n grand, la variable Z suit (approximativement) une loi continue bien connue, appelée loi de Gauss.

Exercice 9.1 [IV.171-174] *Au bridge, le jeu compte cinquante-deux cartes, quatre « couleurs » (trèfle, pique, cœur, carreau) pour chacune des treize figures (as, roi, dame, valet, 2, 3, 4, 5, 6, 7, 8, 9, 10). On constitue une « main » en tirant successivement treize cartes. On s'intéresse au nombre de cartes d'un « atout » (c'est-à-dire d'une « couleur ») fixé à l'avance, mettons par exemple, les carreaux.*

1. *On constitue une main comme on vient de l'expliquer. On note X le nombre de cartes de la couleur « carreau ». Montrer que X suit une loi hypergéométrique de paramètres 52, 13, 13. Montrer qu'alors, si $k+1 \leq 13$, on a la formule de récurrence suivante :*

$$p(X = k+1) = \frac{(13-k)^2}{(k+1)(27+k)} p(X = k)$$

   Calculer explicitement $p(X = 0)$.

2. *On réalise 3400 fois le choix d'une main de bridge. Quelle est la fréquence théorique des valeurs 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 à 13, du nombre de « carreaux » pour ces 3400 épreuves ?*

10 Variables aléatoires réelles et lois continues

Remarque. Si X est une variable continue, par exemple une variable prenant toutes les valeurs possibles dans un intervalle de $\mathbb{R}$, les événements du type ($X =$ constante) sont tous de probabilité nulle. Pour les variables continues, on s'intéressera donc davantage à des événements du type :

$$X \leq \text{constante}$$

ou bien :

$$X \in [x_1, x_2]$$

En général, la loi de probabilité d'une variable aléatoire réelle est entièrement déterminée par les probabilités des événements de ce type (ce sera le cas dans tous les modèles que nous rencontrerons).

Définition 10.1 *Soit $f : \mathbb{R} \to [0,1]$ une fonction intégrable telle que $\displaystyle\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$. On dit d'une variable aléatoire X réelle qu'elle suit une* loi à densité, *et qu'elle a une* densité de probabilité *f, si*

$$(\forall x_1, x_2 \in \mathbb{R}) \quad p(X \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x$$

Définition 10.2 *Sous les hypothèses précédentes, on appelle* fonction de répartition *de X la fonction*

$$F : t \mapsto p(X \leq t) = \int_{-\infty}^{t} f(x)\mathrm{d}x$$

C'est une primitive de f.

Remarque. Pour Δx petit, on a

$$f(x) \simeq \frac{p(X \in [x, x + \Delta x])}{\Delta x}$$

Cette écriture justifie l'appellation « densité de probabilité ».

La loi uniforme sur $[a, b]$. On dit que X suit une loi uniforme sur $[a, b]$ si la densité de probabilité de X est la fonction

$$\frac{\mathbb{1}_{[a,b]}}{b-a}$$

En particulier, si $x_1, x_2 \in [a, b]$, on a

$$p(X \in [x_1, x_2]) = \frac{x_2 - x_1}{b-a}$$

Si $x_1 < x_2 < a < b$, ou si $a < b < x_1 < x_2$, on a $p(X \in [x_1, x_2]) = 0$.

La loi exponentielle de paramètre θ. Sa densité de probabilité est la fonction

$$x \mapsto \theta e^{-\theta x} \mathbb{1}_{[0,+\infty[}$$

Toute variable X suivant une loi exponentielle vérifie la propriété suivante :

$$(\forall t > s > 0) \quad p(X \geq t \mid X \geq s) = p(X \geq t - s \mid X \geq 0)$$

En vertu de cette propriété, la loi exponentielle est souvent utilisée pour modéliser des "durées de vie sans vieillissement".

La loi normale de paramètres μ, σ. Sa densité de probabilité est la fonction

$$x \mapsto \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

On l'appelle aussi *loi de Gauss*, et on dit que X est une *variable normale* ou *gaussienne*. On vérifiera (cf. exercice 10.1 ci-dessous) que cette fonction est bien une densité car son intégrale est égale à 1.

Définition 10.3 *On dit que deux variables aléatoires réelles X et Y sont* indépendantes *si*

$$(\forall x_1, x_2, y_1, y_2 \in \mathbb{R}) \quad p(X \in [x_1, x_2] \ \wedge \ Y \in [y_1, y_2]) = p(X \in [x_1, x_2]).p(Y \in [y_1, y_2])$$

Définition 10.4 *Soit X une variable aléatoire réelle. On définit alors l'espérance* $\mathrm{E}(X)$*, la variance* $\mathrm{V}(X)$*, et l'écart-type* σ_X *:*

$$\mathrm{E}(X) = \int x f(x) \mathrm{d}x$$

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right) = \int (x - \mathrm{E}(X))^2 f(x) \mathrm{d}x$$

$$\sigma_X = \sqrt{V(X)}$$

Propriété 10.1 *L'espérance et la variance ainsi définies vérifient les mêmes propriétés de linéarité, positivité, convexité que pour les variables discrètes. De même, si X et Y sont des variables aléatoires réelles indépendantes, alors* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$*.*

Démonstration. On va seulement montrer cette dernière propriété. Soit donc X et Y deux variables aléatoires réelles indépendantes. On a :

$$\begin{aligned} E(XY) &= \iint xy \; p(X \in [x, x+\mathrm{d}x] \;\wedge\; Y \in [y, y+\mathrm{d}y]) \\ &= \iint xy \; p(X \in [x, x+\mathrm{d}x].p(Y \in [y, y+\mathrm{d}y]) \\ &= \left(\int x \; p(X \in [x, x+\mathrm{d}x]\right) . \left(\int y \; p(Y \in [y, y+\mathrm{d}y]\right) \\ &= \mathrm{E}(X)\mathrm{E}(Y) \end{aligned}$$

Propriété 10.2 *Soit X une variable normale de paramètres μ, σ. Soit $a, b \in \mathbb{R}$ deux constantes. Alors la variable aléatoire $aX + b$ suit une loi normale de paramètres $(|a|\mu + b)$ et $|a|\sigma$.*

Démonstration. (pour $a > 0$)

$$\begin{aligned} p(aX + b \in [x_1, x_2]) &= p\left(X \in \left[\frac{x_1 - b}{a}, \frac{x_2 - b}{a}\right]\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{(x_1-b)/a}^{(x_2-b)/a} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\ &= \frac{1}{a\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(y-(a\mu+b))^2}{2(a\sigma)^2}} \mathrm{d}x \end{aligned}$$

(on a effectué un changement de variable $y = ax + b$ dans l'intégrale).

Conséquence très utile. Soit X une variable normale de paramètres μ, σ. Alors la propriété précédente montre que $(X - \mu)/\sigma$ est une variable normale de paramètres 0 et 1. On dit que c'est une variable normale *centrée réduite*; or il existe des tables de valeurs de la fonction de répartition $\Pi(t)$ des variables normales centrées réduites (cf. tab. 1 p. 28). Ces tables suffisent donc à calculer la probabilité de n'importe quel événement $X \in [x_1, x_2]$ pour toute variable normale X de paramètres μ, σ :

$$\begin{aligned} p(X \in [x_1, x_2]) &= p\left(\frac{X - \mu}{\sigma} \in \left[\frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}\right]\right) \\ &= \Pi\left(\frac{x_2 - \mu}{\sigma}\right) - \Pi\left(\frac{x_1 - \mu}{\sigma}\right) \end{aligned}$$

Exercice 10.1 *1. Soient $r > 0$, $\Delta_r = [-r, r] \times [-r, r]$, et $D_r = \overline{B}(0, r) \subset \mathbb{R}^2$. Montrer que les deux limites suivantes existent et sont égales :*

$$\lim_{r\to+\infty} \iint_{\Delta_r} e^{-x^2-y^2} \mathrm{d}x\mathrm{d}y = \lim_{r\to+\infty} \iint_{D_r} e^{-x^2-y^2} \mathrm{d}x\mathrm{d}y$$

En déduire la valeur de $\displaystyle\int_{-\infty}^{+\infty} e^{-x^2} \mathrm{d}x$.

2. Vérifier que pour $\mu \in \mathbb{R}$ et $\sigma \leq 0$, la fonction

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

est bien la densité d'une loi de probabilité.

Exercice 10.2 *Soit X une variable aléatoire suivant la loi uniforme sur $[a, b]$. Calculer $\mathrm{E}(X)$ et σ_X. Soit Y une variable aléatoire suivant la loi exponentielle de paramètre θ. Montrer que $\mathrm{E}(Y) = \sigma_Y = 1/\theta$. Soit Z une variable normale de paramètres μ, σ. Calculer $\mathrm{E}(Z)$ et σ_Z.*

Exercice 10.3 *On considère la fonction f définie sur $[0, 1]$ par $f(t) = \lambda t^4$, où λ est un réel strictement positif.*

1. *Calculer λ pour que f soit la densité d'une loi de probabilité p définie sur l'intervalle $[0, 1]$*
2. *Calculer $p\left(\left[\frac{1}{5}, \frac{3}{5}\right]\right)$.*

Exercice 10.4 *a est un réel strictement positif.*

1. *Calculer l'intégrale suivante :*

$$I(a) = \int_{-a}^{a} \frac{1}{2} e^{-|x|} dx$$

En déduire $\lim\limits_{a\to+\infty} I(a)$.

2. *Que peut-on en déduire pour la fonction f définie sur $\mathbb{R}$ par $f(x) = \frac{1}{2}e^{-|x|}$?*

Exercice 10.5 [f1] *Soit U une variable aléatoire uniforme sur $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Quelle est la loi de $\tan(U)$? Que vaut $E(|\tan(U)|)$?*

Exercice 10.6 [f1] *Soient X et Y deux variables aléatoires indépendantes de loi exponentielles de paramètre λ. Calculer la loi de $X + Y$.*

Exercice 10.7 [f1] *Soient X et Y deux variables aléatoires indépendantes, on suppose que X suit une loi de densité f et que la loi de Y est portée par $\mathbb{N}$. Calculer la loi de $X + Y$ et de XY. Cas particulier : la loi de X est uniforme sur $[0, 1]$ et la loi de Y est uniforme sur $\{0, 1, ..., n\}$.*

Exercice 10.8 (*paradoxe de Bertrand*) *Soit un triangle équilatéral inscrit dans un cercle. On va tracer une corde « au hasard » dans le cercle. Pour ce faire, on utilise l'une quelconque des trois constructions suivantes.*

1. *On choisit au hasard sur la circonférence du cercle une extrémité de la corde puis, encore au hasard, sa direction. Quelle est la probabilité que la longueur de la corde soit supérieure à celle du côté du triangle équilatéral inscrit ?*

2. *On choisit au hasard la direction de la corde, on trace alors le diamètre perpendiculaire à cette direction, puis, au hasard, le milieu de la corde sur ce diamètre. Quelle est la probabilité que la longueur de la corde soit supérieure à celle du côté du triangle équilatéral inscrit ?*
3. *On choisit au hasard le milieu de la corde sur la surface du cercle. Quelle est la probabilité que la longueur de la corde soit supérieure à celle du côté du triangle équilatéral inscrit ?*

Que répondriez-vous si l'on vous demandait, sans plus de précision, la probabilité que la longueur d'une corde tracée « au hasard » soit supérieure à celle du côté du triangle équilatéral inscrit ?

Exercice 10.9 *On considère la variable aléatoire X égale au poids d'un nourrisson. On suppose que X suit une loi normale de moyenne $m = 3,4$ kg et d'écart-type $\sigma = 0,3$ kg.*

1. *Quelle est la probabilité qu'un nourrisson pèse plus de 4 kg ?*
2. *Quelle est la probabilité qu'il pèse moins de $3,1$ kg ?*
3. *Quelle est la probabilité qu'il pèse entre $3,1$ kg et $3,7$ kg ?*

Exercice 10.10 *On considère la variable aléatoire X qui suit une loi normale de moyenne m et d'écart-type σ. On sait que $p(X \leq 2) = 0,761115$ et que $p(X \geq 6) = 0,00657$. Déterminer la moyenne et l'écart-type de X.*

Exercice 10.11 *Dans un hôpital général, sur toute l'année 2008,* 50000 *hospitalisations ont été comptabilisées, dont* 30000 *en médecine et* 20000 *en chirurgie. Vous disposez des statistiques descriptives suivantes :*

Hospitalisations	*Moyenne ± Ecart-type*	*Médiane*
Total	$N = 50000$	
Durée de séjour (jours)	5 ± 5	x
Age (années)	42 ± 10	y
Chirurgie	$N_1 = 20000$	
Durée du séjour (jours)	4 ± 4	$2,8$
Age (années)	42 ± 10	z

1. *On suppose que l'âge suit une loi normale.*
   (a) *On rappelle que la valeur médiane d'une variable aléatoire continue de fonction de répartition F est le réel t tel que $F(t) = 0,5$. Dans le tableau ci-dessus, que valent y et z ?*
   (b) *L'identité des malades est codée par un numéro d'anonymat sur 10 chiffres, compris entre 0 et 9, sauf le chiffre des unités qui ne peut prendre la valeur 0. Combien de numéros d'identités différents peuvent être définis ?*
   (c) *Quelle est la proportion de séjours d'hospitalisation en 2008 de sujets de plus de 55 ans ?*
   (d) *Si un malade hospitalisé a plus de 40 ans, quelle est la probabilité qu'il en ait plus de 50 ? (Indice : penser à une probabilité conditionnelle)*
2. *On suppose que la durée de séjour suit une loi exponentielle. On rappelle que, pour une loi exponentielle de paramètre θ, la densité de probabilité est $\theta e^{-\theta x}$, l'espérance $1/\theta$, la variance $1/\theta^2$, et pour $0 \leq a \leq b$ on a :*

$$p(X \in [a, b]) = \int_a^b \theta e^{-\theta x} dx$$

(a) *Tracer l'allure du graphe de la densité de probabilité de la durée de séjour des hospitalisations en 2008.*

(b) *Dans le tableau ci-dessus, que vaut x ?*

(c) *Quelle proportion de séjours en 2008 ont duré plus de 5 jours ?*

(d) *Si un malade est hospitalisé depuis 10 jours, quelle est la probabilité qu'il le soit au moins 5 jours de plus (indice : penser à une probabilité conditionnelle) ?*

(e) *Le Directeur de l'hôpital pense que les malades hospitalisés en chirurgie restent plus longtemps à l'hôpital et qu'ils sont plus âgés que les autres. A-t-il raison ?*

Exercice 10.12 *Une machine fabrique en grande série un certain type de pièces rectangulaires en tôle. On note L la variable aléatoire qui, à toute pièce prélevée au hasard dans la production d'une journée, associe sa largeur. On admet que L suit la loi normale de moyenne $58,11$ et d'écart-type $0,15$.*

1. *Déterminer la probabilité p_1 qu'une pièce prélevée au hasard dans cette production ait une largeur comprise entre $57,90$ et $58,30$. Arrondir à 10^{-4}.*
2. *Une pièce a une largeur acceptable lorsque celle-ci est supérieure à $57,90$ (les pièces trop larges pouvant être recoupées). Déterminer la probabilité p_2 qu'une pièce prélevée au hasard dans cette production ait une largeur acceptable. Arrondir à 10^{-3}.*

11 Fonction caractéristique et analyse de Fourier

Rappel. Soit $f : \mathbb{R} \to \mathbb{C}$ une fonction L^1, on peut alors définir sa *transformée de Fourier* :

$$\begin{array}{rccc} \hat{f}: & \mathbb{R} & \longrightarrow & \mathbb{C} \\ & \xi & \longmapsto & \displaystyle\int_{\mathbb{R}} f(x)\exp(-i\xi x)\mathrm{d}x \end{array}$$

Propriété 11.1 *Si $f \in L^1(\mathbb{R})$, alors $\hat{f}$ est une fonction continue.*

Théorème 11.1 (Théorème d'inversion de Fourier) *Soit $f \in L^1(\mathbb{R})$. Supposons que $\hat{f} \in L^1(\mathbb{R})$. Alors la fonction f est presque partout égale à $x \mapsto \hat{\hat{f}}(-x)/(2\pi)$.*

Voir le cours d'analyse de Fourier pour les démonstrations. La démonstration du théorème d'inversion dépend du lemme suivant, qu'il faut ici rappeler car il est en rapport avec la loi normale :

Lemme 11.1 *Soit $f(x) = \exp\left(-\dfrac{x^2}{2}\right)$. Alors :*

$$\hat{f}(\xi) = (2\pi)^{1/2}\exp\left(-\frac{\xi^2}{2}\right).$$

Définition 11.1 *Soit X une variable aléatoire réelle à densité f, alors $f \in L^1(\mathbb{R})$ et on peut définir la* fonction caractéristique *de X :*

$$t \mapsto \Phi_X(t) = \mathrm{E}(\exp(-itX)) = \hat{f}(t)$$

$$\Pi(t) = p(X \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx, \qquad \Pi(-t) = 1 - \Pi(t)$$

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLE 1 – Loi normale centrée réduite

Exemple : si X suit une loi normale de paramètres (μ, σ), on a

$$\Phi_X(t) = e^{-it\mu - \frac{\sigma^2 t^2}{2}}$$

Propriété 11.2 *La fonction caractéristique de X vérifie les propriétés élémentaires suivantes :*

1. $\Phi_X(0) = 1$
2. $(\forall t) \quad |\Phi_X(t)| \leq 1$
3. $\Phi_{-X}(t) = \Phi_X(-t) = \overline{\Phi_X(t)}$
4. *Soit $Y = aX + b$. Alors $\Phi_Y(t) = \exp(-itb)\Phi_X(at)$.*
5. *Si X et Y sont deux variables aléatoires réelles indépendantes, alors $\Phi_{X+Y} = \Phi_X \Phi_Y$.*
6. *Si $\mathrm{E}(|X|^k) < +\infty$, alors Φ_X est de classe C^k et pour tout $j \leq k$, on a $\Phi_X^{(j)}(0) = i^j \mathrm{E}(X^j)$.*

Démonstration. Pour la deuxième :

$$(\forall t) \quad |\Phi_X(t)| \leq \mathrm{E}(|\exp(-itX)|) = 1$$

Pour la quatrième :

$$\Phi_Y(t) = \mathrm{E}(\exp(-it(aX + b))) = \exp(-itb)\mathrm{E}(\exp(-i(at)X))$$

Pour la cinquième :

$$\mathrm{E}(\exp(-it(X + Y))) = \mathrm{E}(\exp(-itX)\exp(-itY)) = \mathrm{E}(\exp(-itX))\mathrm{E}(\exp(-itY))$$

Remarque. En fait, même si $\hat{f}$ n'est pas $L^1(\mathbb{R})$, la donnée de $\hat{f}$ suffit à reconstruire f (voir démonstration du théorème d'inversion). Ainsi l'application $f \mapsto \hat{f}$ est injective, et on a :

Théorème 11.2 *Soit X et Y deux variables aléatoires réelles telles que $\Phi_X = \Phi_Y$. Alors X et Y suivent la même loi, c'est-à-dire que $(\forall\ a, b \in \mathbb{R}) \quad p(X \in [a, b]) = p(Y \in [a, b])$.*

Application. Soit X et Y deux variables indépendantes telles que X suit une loi normale de paramètres (μ_X, σ_X), et Y une loi normale de paramètres (μ_Y, σ_Y). Alors

$$\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t) = \exp\left(-it(\mu_X + \mu_Y) - \frac{(\sigma_X^2 + \sigma_Y^2)t^2}{2}\right)$$

Grâce au théorème précédent, on en conclut que $X + Y$ suit une loi normale de paramètres $(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$. On montrerait de même que $X - Y$ suit une loi normale de paramètres $(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$.

12 Le Théorème Central Limite

Théorème 12.1 *Soit X une variable aléatoire réelle d'espérance m et d'écart-type σ finis. Soit $(X_n)_{n\geq 1}$ une suite de variables indépendantes, chacune de même loi que X. On pose*

$$S_n = X_1 + ... + X_n$$

$$Z_n = \frac{S_n - nm}{\sigma\sqrt{n}}$$

Alors Z_n converge vers une variable Z normale centrée réduite, au sens où :

$$(\forall\ a, b \in \mathbb{R}) \quad \lim_{n\to+\infty} p(Z_n \in [a, b]) = p(Z \in [a, b])$$

Démonstration. Pour travailler avec la somme S_n, il est commode de passer par les fonctions caractéristiques :

$$\begin{aligned}
\Phi_{Z_n}(t) &= \mathrm{E}(\exp(itZ_n)) \\
&= \mathrm{E}\left(\exp\left(i\frac{t}{\sigma\sqrt{n}}(S_n - nm)\right)\right) \\
&= \prod_{j=1}^{n} \mathrm{E}\left(\exp\left(i\frac{t}{\sigma\sqrt{n}}(X_j - m)\right)\right) \\
&= \left(\Phi_{X-m}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n \\
&= \exp\left(n\ln\left(\Phi_{X-m}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\right)
\end{aligned}$$

On applique à Φ_{X-m} la formule de Taylor en 0 :

$$\begin{aligned}
\Phi_{X-m}(u) &= 1 + \Phi'_{X-m}(0)u + \frac{1}{2}\Phi''_{X-m}(0)u^2 + o(u^2) \\
&= 1 + i\mathrm{E}(X-m)u + \frac{1}{2}i^2\mathrm{E}((X-m)^2)u^2 + o(u^2) \\
&= 1 - \frac{\sigma^2u^2}{2} + o(u^2)
\end{aligned}$$

$$\Phi_{Z_n}(t) = \exp\left(n\ln\left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)\right)$$

Alors :

$$\lim_{n\to+\infty} \Phi_{Z_n}(t) = \exp\left(-\frac{t^2}{2}\right)$$

or cette fonction est la fonction caractéristique d'une variable normale centrée réduite Z. (Resterait à montrer que la convergence des fonctions caractéristiques entraîne la convergence des variables aléatoires.)

Remarque. Comme on l'avait annoncé, on peut à présent préciser la loi des grands nombres. Le Théorème Central Limite montre en effet que, pour n grand, la variable aléatoire $\dfrac{S_n}{n}$ suit approximativement une loi normale de paramètres $\mathrm{E}(X)$ et $\dfrac{\sigma}{\sqrt{n}}$.

Exemple. Soit X une variable de Bernoulli de paramètre $\dfrac{1}{2}$, et $n = 100$. S_{100} suit alors une loi binomiale de paramètres $\left(100, \dfrac{1}{2}\right)$, mais si l'on préfère utiliser une table de loi normale plutôt que d'avoir à calculer les factorielles qui interviennent dans l'expression de la loi binomiale, on utilise le Théorème Central Limite : la variable $Y = \dfrac{S_{100}}{100}$ suit approximativement une loi normale de paramètres $\left(\dfrac{1}{2}, \dfrac{1}{20}\right)$.

Critères de validité des approximations d'une loi binomiale. En pratique, on s'autorise à approcher une loi binomiale de paramètres (n, p) par une loi normale de paramètres $\left(np, \sqrt{np(1-p)}\right)$ dès lors que

$$n > 50 \quad \text{et} \quad np(1-p) > 10.$$

Alternativement, si p est petit, on peut approcher cette loi binomiale par une loi de Poisson de paramètre $\lambda = np$ dés lors que

$$n > 30, \quad p < 0,1 \quad \text{et} \quad np < 10.$$

13 Statistiques descriptives

13.1 Les données statistiques

Les statistiques consistent à

- recueillir les valeurs d'une ou plusieurs variables pour chaque individu d'une population ou d'un échantillon d'une population
- décrire ces données au moyen de tableaux, de graphes, et d'« estimateurs » comme la moyenne ou l'écart-type
- faire des hypothèses quant à la loi de probabilité de ces variables, et les tester au vu des données

Soit donc une population donnée de taille N, et une variable aléatoire X décrivant l'état de chaque individu de cette population (par exemple sa taille en centimètres). La collecte des données consiste à répéter la mesure de X pour chaque individu d'une échantillon de taille $n \leq N$. Si l'échantillon est le résultat d'un tirage aléatoire au sein de la population entière, et que N est suffisamment grand (pour que le tirage puisse être conçu comme un « tirage avec remise »), les données statistiques consistent alors en une suite de n variables aléatoires $X_1, X_2, ..., X_n$ indépendantes et de même loi que X. Pour toute valeur x de X, on appelle :

$$\text{fréquence (absolue) de } x = \text{Card}\,\{i \in [\![1, n]\!] \mid X_i = x\}$$

$$\text{fréquence relative de } x = \frac{\text{Card}\,\{i \in [\![1, n]\!] \mid X_i = x\}}{n}$$

$$\text{fréquence cumulée de } x = \text{Card}\,\{i \in [\![1, n]\!] \mid X_i \leq x\}$$

13.2 Représentations tabulaires

13.2.1 Liste des n mesures

La manière la plus simple de représenter les données collectées est d'écrire les résultats des n mesures (dans un ordre quelconque) :

$$X_1, X_2, ..., X_n$$

13.2.2 Série statistique

Lors d'une telle collecte de données, il est possible que plusieurs variables parmi les $X_1, X_2, ..., X_n$ prennent des valeurs égales. Dans tout ce qui suit, on utilisera les lettres miniscules $x_1, x_2, ..., x_k$ pour désigner les k valeurs *distinctes* ($k \leq n$) que l'on a mesurées. On peut alors organiser les données sous forme d'un tableau associant à chaque valeur distincte x_i ($1 \leq i \leq k$) sa fréquence f_i :

$$f_i = \frac{\text{Card}\,\{j \in [\![1, n]\!] \mid X_j = x_i\}}{n}$$

La liste des couples (x_i, f_i) s'appelle une *série statistique*. Si de plus les x_i sont rangés par ordre croissant ($i < j \Rightarrow x_i < x_j$), on en déduit aisément les fréquences cumulées

$$\sum_{m=1}^{i} f_m$$

On obtient ainsi le tableau suivant :

x_1	f_1	f_1
x_2	f_2	$f_1 + f_2$
$\vdots$		
x_k	f_k	$f_1 + f_2 + ... + f_k$

13.2.3 Tableau de contingence

Si on fait des statistiques concernant *deux* variables X et Y et qu'on s'intéresse aux relations mutuelles de ces deux variables, on peut dresser un *tableau de contingence.*

Exemple. Le site internet Morningstar.com rassemble des données sur les fonds de placement. En particulier, il recense une centaine de fonds européens en actions et donne pour chacun une notation (de une étoile `*` à cinq étoiles `*****`), une variable risque (`Low` < `Below Average` < `Average` < `Above Average` < `High`), et un taux annuel (pourcentage). On a copié ces données dans un fichier dont voici quelques lignes :

```
***,Low,1.69
****,Low,2.53
****,Above Average,3.16
****,Average,3.31
*****,Above Average,-1
```

Cherchant à comprendre comment Morningstar note les fonds, on veut étudier le rapport entre la notation et la variable risque. Le code C++ tab. 2 p. 34 dresse alors le tableau de contingence suivant :

	Above Average	Average	Below Average	High	Low
*	4	0	0	6	0
**	1	11	4	2	0
***	2	17	15	1	2
****	9	2	8	2	6
*****	4	2	0	0	4

13.3 Représentations graphiques

13.3.1 Tracé en bâtons

Exemple. Population étudiée : 224414 familles nombreuses qui ont bénéficié d'allocations familiales en 1928. Variable X étudiée = nombre d'enfants par famille bénéficiaire.

2	$7,73\%$
3	$5,11\%$
4	$53,7\%$
5	$21,75\%$
6	8%
7	$2,63\%$
8	$0,76\%$
$X \geq 9$	$0,27\%$

Cf. le *tracé en bâtons* représentant ces données, fig. 3 p. 35.

```
#include <map>
#include <set>
#include <iostream>
#include <iomanip>
#include <string>
#include <sstream>
using namespace std;

int main() {
  map<pair<string,string>,int> tableau;
  set<string> lignes, colonnes;
  string s;

  // boucle pour chaque ligne du fichier
  while(!getline(cin,s).eof()) {
      string rating, risk;
      stringstream buf(s);

      // lire les deux premiers champs
      getline(buf, rating, ','); getline(buf, risk,',');

      // les insérer dans l'ensemble des valeurs possibles
      lignes.insert(rating); colonnes.insert(risk);

      // incrémenter une case du tableau de contingence
      tableau[pair<string,string>(rating,risk)]++;
    }

  set<string>::iterator l,c;

  // écrire les têtes de colonnes
  cout << setw(10) << ' ';
  for (c=colonnes.begin(); c!=colonnes.end(); c++)
    cout << setw(15) << *c;
  cout << endl;

  // écrire le reste du tableau de contingence
  for (l=lignes.begin(); l!=lignes.end(); l++) {
      cout << setw(10) << *l;
      for (c=colonnes.begin(); c!=colonnes.end(); c++)
cout << setw(15) << tableau[pair<string,string>(*l,*c)];
      cout << endl;
    }

  return(0);
}
```

TABLE 2 – Code en C++ qui utilise la *Standard Template Library* (STL) pour produire un tableau de contingence

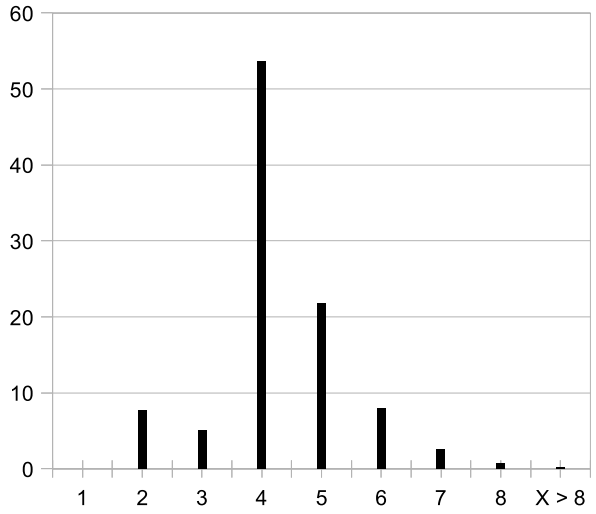


FIGURE 3 – Tracé en bâtons du nombre d'enfants par famille bénéficiant d'allocations (OpenOffice.org)

13.3.2 Histogramme

Considérons à nouveau l'exemple du nombre d'enfants d'une famille ayant bénéficié d'allocations. On peut représenter la série statistique par un *histogramme* dont les ordonnées sont proportionnelles aux fréquences. Quand la variable aléatoire est une variable discrète (comme c'est le cas ici), il faut choisir où porter ses valeurs le long de l'axe des abscisses. OpenOffice.org porte chaque valeur de X au centre de la barre de l'histogramme représentant la fréquence de cette valeur, mais on peut choisir une autre convention : comparer la figure 4 p. 37 et la figure 5 p. 38 où l'on a préféré porter chaque valeur de X à droite de la barre.

Remarque On peut représenter aussi bien les fréquences absolues que les fréquences relatives, seule changera l'échelle sur l'axe des ordonnées.

13.3.3 Polygone intégral des fréquences

Cette courbe dont les ordonnées sont les fréquences cumulées est comparable au graphe de la fonction de répartition de X. A condition d'adopter la bonne convention quand on porte les valeurs de X en abscisses de l'histogramme, on peut tracer l'histogramme et le polygone intégral sur la même figure (*cf.* figure 5 p. 38). Comme son nom l'indique, le polygone intégral représente alors l'intégrale des ordonnées de l'histogramme.

13.4 Cas d'une variable à valeurs nombreuses groupées par classes

Quand une variable a de trop nombreuses valeurs possibles (par exemple s'il s'agit d'une grandeur continue), les tableaux et les histogrammes la décrivant deviennent trop grands, et il est alors plus commode de grouper ces valeurs par classes.

Exemple. La taille des individus d'une population donnée peut être mesurée en centimètres, en prenant la partie entière (par défaut) de la valeur de X trouvée. On prendra donc des classes

```
#include <plotter.h>
#include <string.h>

string X[9]={"1","2","3","4","5","6","7","8","9"};
float f[9]={0,7.73,5.11,53.7,21.75,8,2.63,0.76,0.27};

PSPlotter plotter(cin,cout,cerr);

int main()
{
  plotter.openpl();
  plotter.fspace(-100, -100, 1000, 1100);

  // axe des abscisses
  plotter.fline(0,0,900,0);
  for (int i=0; i<10; i++)
    plotter.fline(i*100, 4, i*100, -4);

  // valeurs de X
  for (int i=0; i<9; i++)
    {
      plotter.fmove((i+1)*100, -7);
      plotter.alabel('c','t',X[i].c_str());
    }

  // histogramme
  for (int i=0; i<9; i++)
    plotter.fbox(i*100, 0, (i+1)*100, f[i]*10.0);

  // polygone intégral des fréquences
  float c=0;
  plotter.linemod("longdashed");
  plotter.fmove(0,0);
  for (int i=0; i<9; i++)
    {
      c+=f[i];
      plotter.fcont((i+1)*100, c*10.0);
    }

  plotter.closepl();
  return 0;
}
```

TABLE 3 – Un histogramme et son polygone intégral : code en C++ avec GNU `libplotter` (X = nombre d'enfants d'une famille bénéficiant d'allocations)

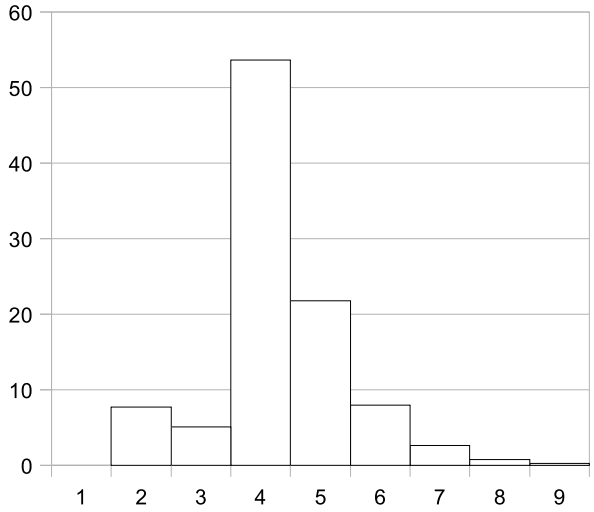


FIGURE 4 – Histogramme (OpenOffice.org)

de 1 cm, par exemple celle composée de tous les individus pour lesquels on a 170 cm $\leq X <$ 171 cm. On dira que l'*intervalle de classe* est de longueur 1 cm. Mais si, pour des calculs ultérieurs on trouve le nombre des classes trop élevé on peut les réunir en classes plus grandes, par groupes de cinq par exemple. Tel est le cas dans la statistique suivante portant sur $n = 334000$ individus, avec des intervalles de classes de longueur 5 cm, et des fréquences exprimées en milliers d'individus :

Intervalles	Fréquences groupées	Fréquences cumulées
$[145, 150[$	1	1
$[150, 155[$	12	13
$[155, 160[$	30	43
$[160, 165[$	90	133
$[165, 170[$	98	231
$[170, 175[$	66	297
$[175, 180[$	27	324
$[180, 185[$	7	331
$[185, 190[$	2	333
$[190, 195[$	1	334

Plus généralement, notons $[\sigma_i, \sigma_{i+1}[$ les intervalles de classes. Les *fréquences groupées* sont alors les f_i :

$$f_i = \text{Card}\,\{j \in [\![1, n]\!] \mid X_j \in [\sigma_i, \sigma_{i+1}[\}$$

Pour de telles données, on n'a plus l'ambiguïté du choix des graduations en abscisses de l'histogramme : la base de chaque rectangle représente l'intervalle de sa classe.

Attention : les fréquences sont-elle représentées par les *hauteurs* des rectangles de l'histogramme, ou bien par les *aires* de ces rectangles ? Si les intervalles des classes sont tous de même longueur, c'est une simple affaire d'échelle en ordonnée. Mais sinon, ces deux choix ne conduisent pas à des représentations semblables. Si l'on choisit de représenter les fréquences par l'aire des rectangles, l'histogramme ressemblera à la courbe représentative de la densité de probabilité de X.

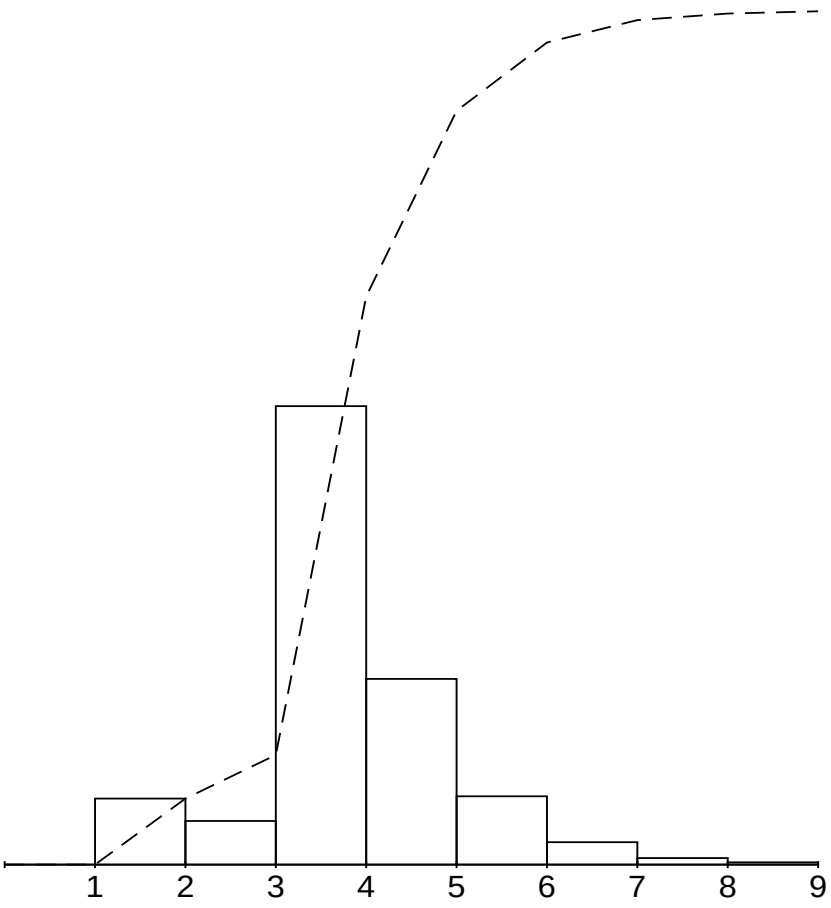


FIGURE 5 – Histogramme et polygone intégral (*cf.* code C++ tab. 3 p. 36)

13.5 La moyenne

La *moyenne* d'une série statistique est le nombre

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{j=1}^{k} f_j x_j$$

Rappel. Ce nombre est une variable aléatoire, et la loi des grands nombres montre qu'il s'agit d'une estimation de $\mathrm{E}(X)$.

Cas de données groupées par classes : on ne connaît que les fréquences groupées et non chaque mesure individuelle. On peut alors utiliser l'approximation suivante :

$$m \simeq \sum_i f_i \frac{\sigma_i + \sigma_{i+1}}{2}$$

Essayons d'évaluer l'erreur ainsi commise. On a :

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} X_j &= \frac{1}{n}\sum_i \left(\sum_{\substack{j \in [\![1,n]\!] \\ X_j \in [\sigma_i, \sigma_{i+1}[}} X_j \right) \\ &= \frac{1}{n}\sum_i n_i m_i \\ &= \sum_i f_i m_i\end{aligned}$$

où

$$\begin{cases} n_i = \mathrm{Card}\,\{j \in [\![1,n]\!] \mid X_j \in [\sigma_i, \sigma_{i+1}[\} \\ m_i = \dfrac{1}{n_i} \displaystyle\sum_{\substack{j \in [\![1,n]\!] \\ X_j \in [\sigma_i, \sigma_{i+1}[}} X_j \end{cases}$$

L'erreur commise en utilisant l'approximation ci-dessus est donc :

$$\sum_i f_i \left(m_i - \frac{\sigma_i + \sigma_{i+1}}{2} \right)$$

Dans le cas où les données sont réparties de manière à peu près symétrique autour de leur moyenne (l'histogramme ressemble à une courbe en cloche), les termes positifs compensent alors assez bien les termes négatifs dans cette somme, et l'approximation est correcte.

13.6 Médiane et quantiles

Soit $p \leq q$ deux entiers positifs. Tout η tel qu'on ait à la fois

$$\begin{cases} \dfrac{1}{n}\mathrm{Card}\{j \in [\![1,n]\!] \mid X_j \leq \eta\} \geq \dfrac{p}{q} \\ \dfrac{1}{n}\mathrm{Card}\{j \in [\![1,n]\!] \mid X_j \geq \eta\} \geq 1 - \dfrac{p}{q} \end{cases}$$

est un *p-ième q-quantile* de la série statistique. Pour $p/q = 1/2$, on dit que η est une *médiane*. Les valeurs de η pour $p/q = 1/4$ et $3/4$ s'appellent des *quartiles*.

Exemple On interroge huit personnes sur le nombre d'emplois qu'ils ont exercés dans le passé. On range ces valeurs par ordre croissant afin d'en déterminer la médiane : 1, 1, 2, 2, 3, 4, 4, 4. Toute valeur dans l'intervalle $[2, 3]$ est alors une valeur médiane. Cela n'aurait pas beaucoup de sens de choisir arbitrairement l'une quelconque de ces valeurs et de dire par exemple que 2,5 est *la* médiane. En effet le nombre d'emploi exercés est une variable discrète à valeurs entières. Alors on dit que $[2, 3]$ est le *segment médian*.

En pratique Soit $X_1, X_2, ..., X_n$ nos n variables aléatoires. Trions-les par ordre croissant :

$$X_{i_1} \leq X_{i_2} \leq X_{i_3} \leq ... \leq X_{i_n}$$

Notons $h = \dfrac{pn}{q}$, et $\mathrm{E}(h)$ sa partie entière (à ne pas confondre avec le E de l'espérance...). De deux choses l'une :

- si $h \in \mathbb{N}$, alors les p-ème q-quantiles sont les éléments de l'intervalle $[X_{i_h}, X_{i_{h+1}}]$;
- sinon, l'unique p-ème q-quantile est le nombre $X_{i_{\mathrm{E}(h)+1}}$.

En effet, pour j donné, essayons de répondre à la question suivante : X_{i_j} est-il un p-ème q-quantile ? On a :

$$\begin{cases} \dfrac{1}{n}\mathrm{Card}\{k \in [\![1, n]\!] \mid X_k \leq X_{i_j}\} \geq \dfrac{j}{n} \\ \dfrac{1}{n}\mathrm{Card}\{k \in [\![1, n]\!] \mid X_k \geq X_{i_j}\} \geq \dfrac{n-j+1}{n} \end{cases}$$

X_{i_j} est donc un p-ème q-quantile dès que $h \leq j \leq h + 1$.

Réciproquement, soit j le plus petit entier vérifiant $h \leq j \leq h + 1$. Si $X_\ell < X_{i_j}$, on aura

$$\mathrm{Card}\{k \in [\![1, n]\!] \mid X_k \leq X_\ell\} < j$$

et donc même (comme ce sont des nombres entiers) :

$$\mathrm{Card}\{k \in [\![1, n]\!] \mid X_k \leq X_\ell\} < h$$

Alors X_ℓ ne peut être un p-ème q-quantile.

Soit j le plus grand entier vérifiant $h \leq j \leq h + 1$. Si $X_\ell > X_{i_j}$, on aura

$$\mathrm{Card}\{k \in [\![1, n]\!] \mid X_k \geq X_\ell\} < n - j + 1$$

et donc même (comme ce sont des nombres entiers) :

$$\mathrm{Card}\{k \in [\![1, n]\!] \mid X_k \geq X_\ell\} < n - h$$

Alors X_ℓ ne peut être un p-ème q-quantile.

Exemple Morningstar a donné la notation « une étoile » (∗) à dix fonds de placement européens en actions (*cf.* 13.2.3 p. 33). Leurs taux annuels sont -3,54 ; -3,15 ; -3 ; -2,8 ; -2,8 ; -2,7 ; 5,08 ; 5,38 ; 5,46 ; 5,46. Le premier quartile ($p/q = 1/4$) de ces taux est donc (-3).

Larges échantillons Si n est grand, les valeurs de X seront assez proches et on aura

$$X_{i_{\mathrm{E}(h)}} = X_{i_{\mathrm{E}(h)+1}}$$

Dans ce cas, on ne perd pas grand chose à dire que, quelque soit $h = \dfrac{pn}{q}$, « le p-ème q-quantile » est $X_{\mathrm{E}(h)+1}$.

Plus généralement S'il faut choisir un p-ème q-quantile, le choix reste purement conventionnel. Une des conventions possibles consiste à prendre le milieu de l'intervalle $[X_{i_h}, X_{i_{h+1}}]$ quand $h \in \mathbb{N}$, et $X_{i_{\mathrm{E}(h)+1}}$ sinon. Ce choix peut d'ailleurs s'écrire ainsi :

$$\frac{X_{i_{\mathrm{E}(h)+1}} + X_{i_{-\mathrm{E}(-h)}}}{2}$$

Les tableurs et les logiciels de statistiques n'ont pas tous adopté une même convention.

Population groupée par classes Dans le cas d'une population groupée par classes, on peut supposer que n est grand. Rangeons ces n individus dans l'ordre croissant des valeurs de la variable étudiée ; le p-ème q-quantile est le h-ème individu, et il s'agit de savoir à quelle classe j il appartient. Pour j donné, les classes 1 à $(j-1)$ comptent $n(f_1 + ...f_{j-1})$ individus, et la j-ème classe en compte nf_j. Il faut donc que

$$n(f_1 + ... + f_{j-1}) \leq h < n(f_1 + ... + f_{j-1} + f_j)$$

Entre σ_j et σ_{j+1}, on suppose les valeurs situées dans la j-ème classe réparties uniformément : le $n(f_1 + ... + f_{j-1})$-ème individu prenant la valeur σ_j, et le $n(f_1 + ... + f_{j-1} + f_j)$-ème prenant la valeur σ_{j+1}, le h-ème individu prendra la valeur suivante obtenue par interpolation linéaire entre ces deux-là :

$$\sigma_j + \frac{h - n(f_1 + ... + f_{j-1})}{nf_j}(\sigma_{j+1} - \sigma_j)$$

C'est le p-ème q-quantile. On procèderait ainsi pour l'exemple du 13.4 p. 35.

Boite à moustaches ou *boxplot*. On représente souvent sur un même graphique le minimum et le maximum, la médiane, et les deux quartiles d'une série statistique. *Cf.* figure 6 p. 42 où on a tracé une boite à moustaches pour chacune des catégories (de * à *****) du classement Morningstar des fonds européens en actions.

13.7 Variance et écart-type

Soit m la moyenne d'une série statistique. Alors sa *variance* s^2 est définie par :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 = \sum_{j=1}^{k} f_j(x_j - m)^2$$

Cas de données groupées par classes. Notons s_i^2 la variance au sein de la i-ème classe (*variance intra-classe*) :

$$s_i^2 = \frac{1}{n_i} \sum_{\substack{j \in [\![1,n]\!] \\ X_j \in [\sigma_i, \sigma_{i+1}[}} (X_j - m_i)^2$$

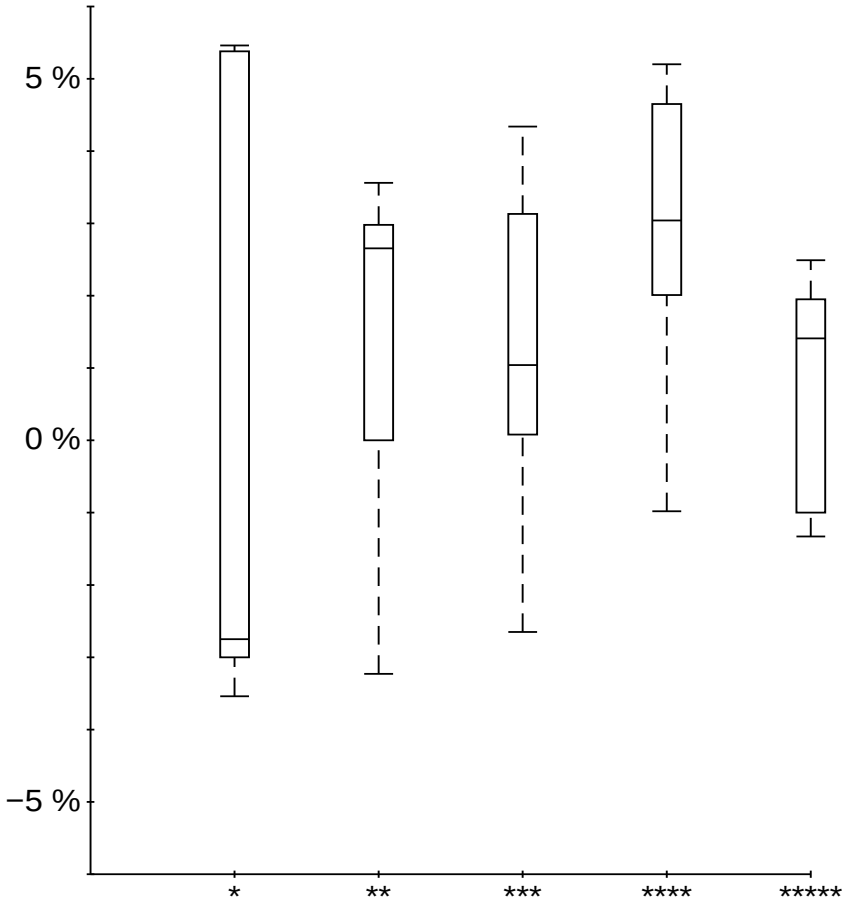


FIGURE 6 – Boites à moustaches des taux annuels des fonds européens en actions pour chaque catégorie de notation Morningstar.

La moyenne de ces variances intra-classes est

$$\sum_i f_i s_i^2$$

Remarquons que la variance de X est égale à la variance de $X-m$, et de même au sein de chaque classe, donc :

$$\begin{aligned} s_i^2 &= \frac{1}{n_i} \sum_{\substack{j \in [\![1,n]\!] \\ X_j \in [\sigma_i, \sigma_{i+1}[}} (X_j - m)^2 - \left(\frac{1}{n_i} \sum_{\substack{j \in [\![1,n]\!] \\ X_j \in [\sigma_i, \sigma_{i+1}[}} (X_j - m) \right)^2 \\ &= \frac{1}{n_i} \sum (X_j - m)^2 - (m_i - m)^2 \end{aligned}$$

Or

$$s^2 = \frac{1}{n} \sum_{j \in [\![1,n]\!]} (X_j - m)^2 = \frac{1}{n} \sum_i \left(\sum_{\substack{j \in [\![1,n]\!] \\ X_j \in [\sigma_i, \sigma_{i+1}[}} (X_j - m)^2 \right)$$

donc

$$\sum_i f_i s_i^2 = s^2 - \sum_i f_i (m_i - m)^2$$

On appelle le dernier terme *variance inter-classes*. On a ainsi décomposé la variance s^2 en somme de la variance inter-classes et de la moyenne des variances intra-classes :

$$s^2 = \sum_i f_i (m_i - m)^2 + \sum_i f_i s_i^2$$

En pratique, les m_i ne sont pas connues et on calcule s^2 de manière approchée en faisant

$$s^2 \simeq \sum_i f_i \left(\frac{\sigma_i + \sigma_{i+1}}{2} - m \right)^2$$

L'erreur commise est alors :

$$\sum_i f_i \left((m_i - m)^2 - \left(\frac{\sigma_i + \sigma_{i+1}}{2} - m \right)^2 \right) + \sum_i f_i s_i^2$$

Estimateur sans biais de la variance. Posons $Y = \frac{1}{n} \sum_{j=1}^{n} X_j$, et $V = \frac{1}{n} \sum_{j=1}^{n} (X_j - Y)^2$. On sait (loi des grands nombres) que pour n grand $Y \simeq \mathrm{E}(X)$. Plus exactement, $\mathrm{E}(Y) = \mathrm{E}(X)$ et $\lim_{n \to +\infty} \mathrm{V}(Y) = 0$. En revanche, on montre que

$$\mathrm{E}(V) = \frac{n-1}{n} \mathrm{V}(X)$$

(*cf.* exercice 13.9 p. 46). On dit que l'estimateur V de la variance de X présente un *biais*. Pour n petit, ce biais n'est pas négligeable et on préfère donc utiliser un « estimateur sans biais » : on pose

$$s_e^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - m)^2$$

On appelle s_e^2 l'*estimateur sans biais de la variance*.

13.8 Exercices

Exercice 13.1 *Dans une entreprise, on a demandé à huit analystes programmeurs d'évaluer le taux de réutilisation du code source quand ils développent un nouveau logiciel. Les données suivantes expriment le pourcentage, dans la totalité du code d'un logiciel, du code issu de dépots réutilisés :*

$$50 \quad 62,4 \quad 37,5 \quad 75 \quad 45 \quad 47,5 \quad 15 \quad 25$$

Calculez la moyenne, la médiane, l'étendue, la variance, l'écart-type, l'écart moyen d'ordre 1, puis interprétez.

Exercice 13.2 *Le tableau ci-dessous donne la répartition des salariés d'une entreprise suivant leur salaire mensuel moyen :*

Tranches de salaires (euros)	*Effectifs*
$[0, 250[$	10
$[250, 500[$	15
$[500, 750[$	45
$[750, 1000[$	110
$[1000, 1250[$	255
$[1250, 1500[$	150
$[1500, 1750[$	60
$[1750, 2000[$	35
$[2000, 2500[$	20

1. *Représenter graphiquement la série par un histogramme et le polygone intégral des effectifs.*
2. *Donner l'étendue de la série, sa classe modale.*
3. *Quel est le salaire moyen mensuel payé par l'entreprise ?*
4. *Déterminer la médiane de la série.*
5. *Quel est le pourcentage des salariés de l'entreprise ayant un salaire dans l'intervalle* $[750, 1750[$ *?*

Exercice 13.3 *La taille x des individus d'une population fait l'objet de la statistique suivante,*

portant sur $N = 334000$ *individus :*

Classes de tailles (cm)	*Effectifs (en milliers)*	*Effectifs cumulés*
$[145, 150[$	1	1
$[150, 155[$	12	13
$[155, 160[$	30	43
$[160, 165[$	90	133
$[165, 170[$	98	231
$[170, 175[$	66	297
$[175, 180[$	27	324
$[180, 185[$	7	331
$[185, 190[$	2	333
$[190, 195[$	1	334

Représenter graphiquement la série par un histogramme et le polygone intégral des fréquences. Donner son étendue, sa classe modale, sa moyenne, sa médiane, sa variance, son écart-type. Pour calculer la moyenne et la variance de x*, on calculera d'abord celles de*

$$\xi = \frac{x - 170}{5}$$

Exercice 13.4 *On considère les familles nombreuses qui ont bénéficié d'allocations. On les dénombre suivant le nombre* X *d'enfants par famille bénéficiaire, soit (en 1928) :*

X	2	3	4	5	6	7	8	$X \geq 9$
n	17331	11455	120614	48813	17963	5919	1713	606

Calculer les pourcentages de chaque valeur du caractère étudié. Quel est le mode ? la médiane ? la moyenne de X *? son étendue ? sa variance ? son écart-type ? Tracer l'histogramme et le polygone intégral.*

Exercice 13.5 *On note* m *la moyenne,* η *la médiane,* μ *l'écart moyen d'ordre 1, et* s *l'écart-type. Le but de cet exercice est de montrer qu'on a* $|\eta - m| \leq \mu$*. On s'aidera des indications suivantes :*

1. *Prendre* m *pour origine* ($m = 0$)*, soit*

$$\sum f_k x_k = 0.$$

   Dans cette somme, la constribution des termes strictement négatifs est donc égale, en valeur absolue, à celle des termes strictement positifs.

2. *Supposer* $\eta > 0$*, alors dans la somme* $\mu = \sum f_k |x_k|$*, la contribution des* $x_k \geq \eta$ *est au moins égale à* $\eta/2$*. Donc la remarque précédente entraîne que*

$$\mu \geq \frac{\eta}{2} + \frac{\mu}{2}$$

3. *Conclure.*

Exercice 13.6 [To32] *Le but de cet exercice est de montrer que* $\mu \leq s$*.*

1. *Démontrer l'identité de Lagrange valable pour* N *couples quelconques de nombres* (a_i, b_i) *:*

$$\left(\sum_{i=1}^{N} x_i y_i\right)^2 + \sum_{1 \leq i < j \leq N} (x_i y_j - x_j y_i)^2 = \left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{j=1}^{N} y_j^2\right)$$

2. *En prenant* $a_k = \sqrt{f_k}.x_k$ *et* $b_k = \sqrt{f_k}$, *et en tenant compte de* $\sum f_k = 1$, *montrer que :*

$$\left(\sum f_k x_k\right)^2 \leq \sum f_k x_k^2.$$

3. *Appliquer cette inégalité à la variable* $|X - m|$ *et conclure.*

Exercice 13.7 [To32] *Relations entre* m, η, μ *et* s *(moyenne, médiane et écarts). Déduire des exercices précédents que :*

$$|m - \eta| \leq \mu \leq s.$$

Montrer que la double égalité est obtenue si et seulement si la répartition se réduit à deux valeurs x_1 *et* x_2 *(*$f_1 = f_2 = 0,5$*). S'il y a un segment médian, celui-ci est compris dans le segment* $[m - \mu, m + \mu]$, *lui-même compris dans le segment* $[m - s, m + s]$.

Exercice 13.8 [To33] *Montrer que* $\overline{|X - a|}$ *atteint sa valeur minima lorsque a est valeur médiane.*

Indications : *soit* $a > 0$ *quelconque. La moyenne de* $|X - a|$ *est la somme des trois termes* X_-, $X_{0,a}$ *et* X_+ *relatifs respectivement aux trois intervalles* $x \leq 0$, $0 < x < a$ *et* $x \geq a$. *En déduire l'égalité :*

$$\overline{|x - a|} = \overline{|x|} + 2X_{0,a} + a(f(x \leq 0) - f(x > 0)),$$

$f(.)$ *désignant les fréquences relatives des valeurs de* x *indiquées entre parenthèses. Une égalité analogue vaut pour* $a < 0$, *avec, pour troisième terme :*

$$|a|(f(x \geq 0) - f(x < 0))$$

En conclure, $X_{0,a}$ *étant* ≥ 0, *que :*

1. *si* $x = 0$ *est une valeur médiane,* $\overline{|x - a|}$ *prend sa valeur minimum lorsque* $a = 0$ *;*
2. $\overline{|x - a|}$ *prend cette même valeur minimum pour toute valeur médiane a ;*
3. *si* 0 *et a sont deux valeurs médianes,* $f(0 < x < a) = 0$.

13.9 Échantillons issus aléatoirement d'une population finie

Exercice 13.9 [IV.17-19] *Soit une urne contenant des boules portant chacune un numéro (il peut arriver que plusieurs boules portent un même numéro). On fait n tirages successifs* avec remise *; les variables aléatoires* X_1, X_2,..., X_n *dénotent les résultats respectifs de ces tirages. Montrer que ces variables aléatoires sont indépendantes et ont toutes la même loi. On suppose que* $\mathrm{E}(X_1) = 0$. *On considère encore les variables aléatoires suivantes :*

$$Y = \frac{X_1 + ... + X_n}{n}$$

$$V = \frac{1}{n}\sum_1^n (X_i - Y)^2$$

On se propose de calculer l'espérance et la variance de V. *Pour cela, démontrer successivement les formules suivantes.*

$$V = \left(\frac{1}{n}\sum_1^n X_i^2\right) - Y^2$$

$$\mathrm{E}(V) = \frac{(n-1)\mathrm{V}(X_1)}{n}$$

$$n^2V^2 = (X_1^2 + ... + X_n^2)^2 - 2nY^2\left(\sum X_i^2\right) + \frac{\left(\sum X_i\right)^4}{n^2}$$

$$\mathrm{E}(X_1^2X_2^2) = \mathrm{V}(X_1)^2$$

$$\mathrm{E}\left(\left(\sum X_i\right)^4\right) = \sum \mathrm{E}(X_i^4) + 6\sum \mathrm{E}(X_i^2X_j^2) = n\mathrm{E}(X_1^4) + 3n(n-1)\mathrm{V}(X_1)^2$$

$$\mathrm{V}(V) = \frac{\mathrm{E}(X_1^4)}{n}\left(1 - \frac{1}{n}\right)^2 - \frac{\mathrm{V}(X_1)^2}{n}\left(1 - \frac{4}{n} + \frac{3}{n^2}\right)$$

Exercice 13.10 *Soit une urne contenant un nombre fini N de boules portant chacune un numéro (il peut arriver que plusieurs boules portent un même numéro, et un numéro peut être négatif). On fait N tirages successifs sans remise ; les variables aléatoires X_1, X_2,..., X_N dénotent les résultats respectifs des N tirages. Montrer que ces variables aléatoires ont toutes la même loi. On suppose que $\mathrm{E}(X_1) = 0$.*

a) Montrer que $X_1 + X_2 + ... + X_N = 0$, puis que

$$\left(\sum_1^N X_i\right)^2 = 0$$

b) En développant, en déduire que

$$N\mathrm{V}(X_1) + N(N-1)\mathrm{E}(X_1X_2) = 0$$

d'où :

$$\mathrm{E}(X_1X_2) = -\frac{\mathrm{V}(X_1)}{N-1}$$

Exercice 13.11 [IV.15] *On se place dans les même conditions que l'exercice précédent.*

a) On fait seulement n tirages ($n \leq N$). On introduit une nouvelle variable aléatoire Y définie par :

$$Y = \frac{X_1 + ... + X_n}{n}$$

Montrer que $n^2\mathrm{V}(Y) = n\mathrm{V}(X_1) + n(n-1)\mathrm{E}(X_1X_2)$, puis que :

$$\mathrm{V}(Y) = \frac{\mathrm{V}(X_1)(N-n)}{n(N-1)}$$

b) On suppose que $2n \leq N$, ce qui permet de faire successivement deux fois n tirages. On note :

$$Z = X_1 + X_2 + ... + X_n$$

$$Z' = X_{n+1} + X_{n+2} + ... + X_{2n}$$

Montrer que

$$\mathrm{cov}(Z, Z') = -\frac{n^2\mathrm{V}(X_1)}{N-1}$$

puis que le rapport $\dfrac{\mathrm{cov}(Z, Z')}{\mathrm{V}(Z)}$ est indépendant de la loi de X_1.

c) Appliquer cet exercice au nombre d'as dans deux « mains » au bridge. Le jeu compte 52 cartes dont quatre as. On constitue une « main » en tirant successivement 13 cartes. On suppose que, pour tout i, la variable aléatoire X_i vaut $48/52$ ou $-4/52$ suivant que la i-ème carte tirée est ou n'est pas un as. Montrer qu'alors on a bien $\mathrm{E}(X_1) = 0$, et que $Z+1$ et $Z'+1$ dénotent le nombre d'as dans chaque main. Calculer ensuite $\mathrm{cov}(Z, Z')$.

14 Statistiques inférentielles

14.1 Intervalles de confiance

14.1.1 Intervalle de confiance d'une moyenne

Comme on l'a vu, la situation rencontrée en statistiques est souvent la suivante : dans une population, on constitue un échantillon de taille n en faisant des tirages avec remise (si n/N est petit, on peut toujour assimiler un tirage aléatoire sans remise à un tirage avec remise). Calculons la moyenne $\dfrac{S_n}{n} = \dfrac{X_1 + X_2 + ... + X_n}{n}$. Alors le théorème central limite montre que pour n grand, $\dfrac{S_n}{n}$ suit approximativement une loi normale de paramètres $\left(\mathrm{E}(X), \dfrac{\sigma_X}{\sqrt{n}}\right)$. En particulier, $\dfrac{S_n}{n}$ est un bon estimateur de $\mathrm{E}(X)$ pour n grand.

On voudrait à présent estimer l'espérance en précisant l'erreur que l'on risque de commettre. C'est-à-dire qu'on voudrait pouvoir affirmer : « j'ai confiance à ...% en le fait que l'espérance $\mathrm{E}(X)$ est située dans l'intervalle $\left[\dfrac{S_n}{n} - ..., \dfrac{S_n}{n} + ...\right]$ ».

Notons Z la variable centrée réduite :

$$Z = \frac{\dfrac{S_n}{n} - \mathrm{E}(X)}{\dfrac{\sigma_X}{\sqrt{n}}}$$

Alors pour tout $z \in \mathbb{R}_+$

$$p\left(\mathrm{E}(X) \in \left[\frac{S_n}{n} - \frac{z\sigma}{\sqrt{n}},\ \frac{S_n}{n} + \frac{z\sigma}{\sqrt{n}}\right]\right) = p(Z \in [-z, z]) = 2\Pi(z) - 1$$

où Π est la fonction de répartition de la loi normale centrée réduite. Les seuils de confiance couramment utilisés sont indiqués dans le tableau suivant avec les valeurs de z correspondantes :

$2\Pi(z) - 1$	90%	95%	99%	99,5%	99,8%
z	1,645	1,96	2,58	2,81	3,08

Par exemple, on peut affirmer : « j'ai confiance à 95 % en le fait que l'espérance $\mathrm{E}(X)$ est située dans l'intervalle $\left[\dfrac{S_n}{n} - \dfrac{1,96\sigma}{\sqrt{n}}, \dfrac{S_n}{n} + \dfrac{1,96\sigma}{\sqrt{n}}\right]$ ».

Exemple. Soit X la variable aléatoire dénotant la taille d'un individu d'une population donnée. On veut estimer $\mathrm{E}(X)$ au moyen d'un échantillon de quarante individus. Il faut d'abord calculer la taille moyenne m et l'écart-type s de cette série statistique. L'intervalle de confiance à 95 % est alors

$$\left[m - \frac{1,96s}{\sqrt{40}}, m + \frac{1,96s}{\sqrt{40}}\right]$$

Remarque. L'*inégalité de Bienaymé-Chebyshev* permet d'écrire des intervalles de confiance sans même avoir recours au théorème central limite, mais il sont beaucoup plus lâches. Voir exercice 14.1 p. 49 et suivants.

14.1.2 Intervalle de confiance d'une fréquence

Soit $x \in X(\Omega)$. On veut estimer la probabilité $p = p(X = x)$. Définissons les variables $Y_1, Y_2, ..., Y_n$ telles que

$$Y_i = \begin{cases} 1 \text{ si } X_i = x \\ 0 \text{ sinon} \end{cases}$$

Alors les Y_i sont des variables de Bernoulli d'espérance $\mathrm{E}(Y_i) = p$ et d'écart-type $\sigma_{Y_i} = \sqrt{p(1-p)}$. La moyenne $m = \dfrac{\sum Y_i}{n}$ n'est autre que la fréquence de x. La loi des grands nombres révèle que cette fréquence est une bonne estimation de p. On peut alors appliquer les résultats de la section précédente aux variables $Y_1, Y_2, ..., Y_n$ pour avoir un intervalle de confiance :

$$p\left(p \in \left[m - z\sqrt{\frac{m(1-m)}{n}}, m - z\sqrt{\frac{m(1-m)}{n}}\right]\right) = 2\Pi(z) - 1$$

14.1.3 Exercices

Exercice 14.1 *Soit X une variable aléatoire réelle positive (et on suppose que $\mathrm{E}(X) > 0$). Soit $\lambda > 0$. On définit la variable aléatoire Y ainsi :*

$$Y(\omega) = \begin{cases} \lambda\mathrm{E}(X) \text{ si } X(\omega) \geq \lambda\mathrm{E}(X) \\ 0 \text{ sinon} \end{cases}$$

Montrer que $\mathrm{E}(Y) \leq \mathrm{E}(X)$, puis que :

$$\mathrm{E}(Y) = p\left(X \geq \lambda\mathrm{E}(X)\right)\lambda\mathrm{E}(X)$$

En déduire que (« inégalité de Markov ») :

$$p\left(X \geq \lambda\mathrm{E}(X)\right) \leq \frac{1}{\lambda}$$

Soit à présent une autre variable aléatoire réelle Z. Utiliser l'inégalité de Markov pour majorer $p\left((Z - \mathrm{E}(Z))^2 \geq \lambda\mathrm{V}(Z)\right)$. En déduire que, pour tout $\epsilon > 0$, on a (« inégalité de Bienaymé-Tchebychev ») :

$$p(|Z - \mathrm{E}(Z)| \geq \epsilon) \leq \frac{\mathrm{V}(Z)}{\epsilon^2}$$

Exercice 14.2 *On donne ici des renseignements sur la distribution de la taille (en inch) des habitants des îles britaniques (on rappelle pour mémoire qu'un inch vaut environ 2,5 mm) :*

	Angleterre	*Ecosse*	*Pays de Galles*	*Irlande*
moyenne	67.31	68.55	66.62	67.68
médiane	67.35	68.48	66.56	67.69
écart-type	2.56	2.50	2.35	2.17
1^{er} quartile	65.55	66.92	65.00	66.39
$2^{ème}$ quartile	69.10	70.04	67.98	69.1
minimum	56.60	59.80	60.60	60.8
maximum	77.4	77.30	74.30	72.3

1. *Représentez chacune de ces quatre distributions par un « box plot ». Que nous apprend ce dessin ?*

2. *Utilisez l'inégalité de Bienaymé Chebishev pour donner, pour chacune des distributions, un intervalle contenant au moins 75 % de la population.*

Exercice 14.3 *Le tableau suivant donne la production laitière (en gallons par semaine) de 4912 vaches (ces données datent de 1922) :*

quantité de lait (gallons par semaine)	*nombre de vaches*
$[7.5, 12.5[$	123
$[12.5, 15.5[$	726
$[15.5, 18.5[$	1636
$[18.5, 21.5[$	1530
$[21.5, 26.5[$	821
$[26.5, 33.5[$	76
Total	4912

1. *Quelle est la population étudiée dans ce tableau, quelle est la variable ?*
2. *Représentez cette distribution par un histogramme.*
3. *A combien estimez vous la quantité totale de lait produite par les 1636 vaches qui produisent entre 15.5 et 18.5 litres de lait par semaine ?*
4. *Calculez la production moyenne de l'ensemble des vaches, ainsi que la production médiane.*
5. *Calculez la variance de la production laitière hebdomadaire des 42912 vaches, ainsi que l'écart-type de cette distribution.*
6. *Utilisez l'inégalité de Bienaymé Chebyshev pour donner un intervalle contenant la production d'au moins 75% des vaches. A l'aide du tableau des données et en utilisant la technique de l'interpolation linéaire, donnez le nombre approximatif de vaches dont la production appartient réellemen à cette intervalle.*

Exercice 14.4 *On a testé la consommation d'essence aux 100 km d'une voiture :*

litres aux 100 km	*nombre d'essais*	*centres des classes*	*fréquences cumulées*
$[9.6, 9.7[$	14		
$[9.7, 9.8[$	25		
$[9.8, 9.9[$	31		
$[9.9, 10[$	17		
$[10, 10.1[$	13		

1. *Compléter le tableau.*
2. *Donner l'étendue de la série.*
3. *Dessiner l'histogramme de la série.*
4. *Calculer la consommation moyenne de la voiture, puis son écart-type.*
5. *Représenter le polygone intégral (courbe des fréquences cumulées) de la série et déterminer graphiquement sa médiane.*
6. *Utiliser l'inégalité de Bienaymé-Chebyshev pour donner un intervalle contenant la consommation d'au moins 75 % des voitures.*
7. *Selon Bienaymé-Chebyshev, quel pourcentage de ces voitures vous attendez-vous à trouver à ± 3 écarts-type de la consommation moyenne ?*

Exercice 14.5 *On considère une population de 1024 fonds communs de placement qui investissent surtout dans de grandes entreprises. On a déterminé que la performance annuelle moyenne de ces fonds est* $\mu = 28,20$ *et que son écart-type est* $\sigma = 6.75$*. Supposons aussi que l'étendue de la performance annuelle va de* $0,3$ *à* $60,3$*, et que les quartiles sont* $23,9$ *et* $32,3$*.*

1. *Si les performances annuelles des fonds ont une répartition qui ressemble à une « courbe en cloche », empiriquement, quel pourcentage de ces fonds vous attendez-vous à trouver à* ± 1 *écart-type de la performance moyenne ? Et à* ± 2 *écarts-type ?*
2. *Selon Bienaymé-Chebyshev, quel pourcentage de ces fonds vous attendez-vous à trouver à* ± 1 *écart-type de la performance moyenne ? À* ± 2 *écarts-type ? À* ± 3 *écarts-type ?*
3. *Selon Bienaymé-Chevyshev, au moins* $93,75\%$ *de ces fonds auront une performance annuelle comprise entre ... et ...*

14.2 Tests d'hypothèses

14.2.1 Un exemple

Pour un âge donné, la taille est distribuée selon une loi normale. La taille des enfants de quatre ans suit une loi normale de paramètres (100 cm, 4 cm). Des biologistes affirment que la taille des enfants de quatre ans *ayant une certaine maladie chronique* est en moyenne 84 cm avec une écart-type de 3,2 cm. Vous tentez de vérifier cette affirmation en faisant des statistiques sur un échantillon de 100 enfants atteint de cette maladie. Vous calculez une taille moyenne de 90 cm, avec un écart-type de 4 cm. Qu'en conclure ?

Notons X la taille d'un enfant malade. En l'absence de toute hypothèse sur cette variable aléatoire, les statistiques nous révèlent que son espérance est proche de 90 cm. On peut même calculer un intervalle de confiance à 95 % :

$$\left[90 - \frac{1,96 \times 4}{\sqrt{100}}, 90 + \frac{1,96 \times 4}{\sqrt{100}}\right] = [89,22; 90,78]$$

Puisque cet intervalle ne contient pas 84 cm, on doute fort de l'affirmation des biologistes. On tendrait plutôt à croire (« à 95 % ») que l'espérance de X est comprise entre 89,22 cm et 90,78 cm.

On peut aussi raisonner de la manière suivante. Faisons l'hypothèse que l'affirmation des biologistes est vraie. Notons (H_0) cette hypothèse :

(H_0) la taille des enfants de 4 ans malades suit une loi normale de paramètres 84 cm et 3,2 cm

Sous cette hypothèse, en adoptant les notations de la section 14.1.1, on aura

$$p\left(\mathrm{E}(X) \in \left[m - \frac{1,96\sigma_X}{\sqrt{100}},\ m + \frac{1,96\sigma_X}{\sqrt{100}}\right]\right) = 95\%$$

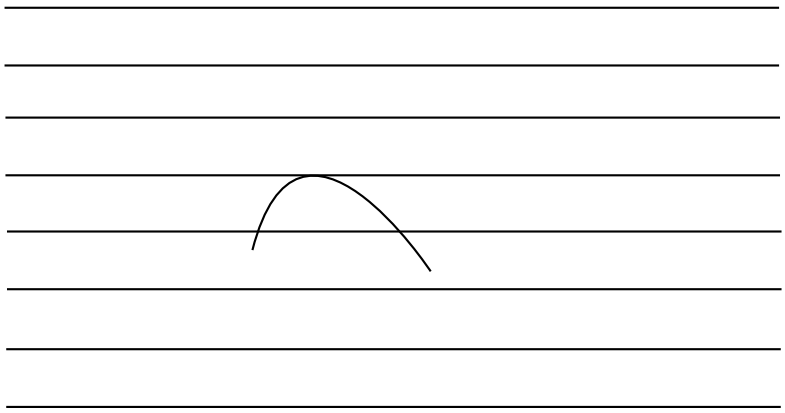
où *a priori* $\mathrm{E}(X) = 84$ cm et $\sigma_X = 3,2$ cm, donc

$$p\left(m \in \left[84 - \frac{1,96 \times 3,2}{\sqrt{100}},\ 84 + \frac{1,96 \times 3,2}{\sqrt{100}}\right]\right)$$
$$= p(m \in [83,37; 84,63]) = 95\%$$

L'événement $m \in [83,37; 84,63]$ a une probabilité de 95 %, et son complémentaire $m \notin [83,37; 84,63]$ une probabilité de 5 %. Nos statistiques fournissent une valeur de m (ici $m = 90$ cm). Nous déciderons de rejeter l'hypothèse (H_0) précisément lorsque $m \notin [83,37; 84,63]$. Sous l'hypothèse que (H_0) est vraie, la probabilité que cet événement se produise et nous conduise à commettre une erreur en rejetant (H_0) est alors seulement 5 %.

14.3 Problème : l'aiguille de Buffon

Question 1 On lance au hasard un arc de courbe $\mathcal{C}$ plan et rigide, une sorte de spaghetti cru et courbe, de forme quelconque et de longueur ℓ, sur un parquet. Les lames de bois de largeur h qui constituent le parquet ont des bords rectilignes, parallèles. On note X le nombre de points d'intersection de la courbe $\mathcal{C}$ avec les bords des lames du parquet. Vous avez suivi des cours de statistiques en L1. Choisissez donc une courbe $\mathcal{C}$ particulièrement simple (par exemple un segment de droite), et décrivez comment vous mèneriez une étude statistique de la variable X. Eventuellement, réalisez vous-mêmes l'expérience ; sinon, expliquez comment vous pourriez la réaliser, jusqu'à la collecte des données puis leur analyse.



Exemple : $X = 3$

Question 2 Admettons désormais qu'il existe un espace de probabilité $(\Omega, \mathcal{A}, p)$ décrivant cette expérience. Dans un premier temps, supposons que la courbe $\mathcal{C}$ est un segment de droite [AB]. Soit I le milieu de [AB]. Soit X_1 le nombre de points d'intersection du segment [AI] avec les bords des lames de parquet, et X_2 le nombre de points d'intersection du segment [IB] avec les bords des lames. Comparer $\mathrm{E}(X)$, $\mathrm{E}(X_1)$, $\mathrm{E}(X_2)$.

Question 3 Supposons maintenant que la courbe $\mathcal{C}$ est un polygone équilatéral. Montrer que $\mathrm{E}(X)$ est proportionnel à ℓ :

$$\mathrm{E}(X) = k\ell$$

Généralisation à une courbe $\mathcal{C}$ quelconque de longueur ℓ ? *Indication :* on assimilera la courbe à un polygone équilatéral ayant une infinité de côtés de longueur ds.

Question 4 Supposons maintenant que la courbe $\mathcal{C}$ est un cercle de rayon r. Montrez que, pour certaines valeurs de r, la variable aléatoire X est en fait une constante. Que vaut alors $\mathrm{E}(X)$? En déduire la valeur de k.

Question 5 Faire le lien avec l'étude statistique de la question **1)** : choisir une courbe $\mathcal{C}$ particulièrement simple (par exemple un segment de droite), puis montrer comment estimer statistiquement $\mathrm{E}(X)$, et donc k. En déduire un procédé délivrant une estimation de la constante

π, circonférence du cercle de rayon unité. Réalisez l'algorithme dans un langage de programmation de votre choix.

Indication : il faudra alors réfléchir à ce que signifie « lancer au hasard » pour concevoir un simulateur qui utilise un générateur de nombres aléatoires (fonction *random*). Par exemple, on peut décider que « lancer au hasard » consiste à choisir successivement :

(i) un point au hasard sur le parquet comme origine de la courbe $\mathcal{C}$

(ii) une direction au hasard c'est-à-dire un angle compris entre 0 et 2π

Question 6 On lance à présent n courbes $\mathcal{C}$ identiques. Soit X_i le nombre de points d'intersection de la $i^{\text{ème}}$ courbe avec les bords des lames. Soit Z_n la variable aléatoire définie par :

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(c'est le nombre moyen de points d'intersection). Le théorème *central limit* (que l'on admettra) affirme que pour n grand, la variable Z_n suit approximativement une loi normale de paramètres μ, σ avec

$$\mu = \mathrm{E}(X)$$
$$\sigma = \frac{\sigma_X}{\sqrt{n}}$$

Expliquez comment estimer statistiquement σ_X. Au moyen d'une table de loi normale, trouver un réel positif ε dépendant de n et de σ_X, tel que

$$p\left(Z_n \in [\mathrm{E}(X) - \varepsilon, \mathrm{E}(X) + \varepsilon]\right) \gtrsim 95\ \%$$

Question 7 La question précédente permet d'évaluer l'erreur commise quand on estime $\mathrm{E}(X)$ au moyen de la variable aléatoire Z_n. On peut alors « avoir confiance » (à 95 %) que $\Delta\mathrm{E}(X) = |Z_n - \mathrm{E}(X)| < \varepsilon$. Notons $\Delta\pi$ l'incertitude sur l'estimation de π délivrée par l'algorithme de la question **5)**. Comparez $\dfrac{\Delta\pi}{\pi}$ et $\dfrac{\Delta\mathrm{E}(X)}{\mathrm{E}(X)}$. Combien de lancers faut-il faire pour obtenir une estimation de π à 10^{-4} près ? Perfectionnez votre algorithme afin qu'il fournisse une estimation de π à une précision quelconque passée en argument. *Indication :* il faudra procéder en deux temps, puisque le calcul du nombre de lancers requiert des estimations grossières préalables de $\mathrm{E}(X)$ et σ_X.

Question 8 Si $\mathcal{C}$ est un segment de droite de longueur $\ell < h$, vérifiez la formule trouvée dans les questions **3)** et **4)** en déterminant explicitement la loi de probabilité de la variable X puis en faisant un calcul d'espérance *(question à traiter après le cours sur les lois de couple).*

15 Statistique multivariée

15.1 Corrélation

Nuage de points Soit à étudier deux variables X et Y et un échantillon de n individus. Les données sont les couples

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$$

que l'on peut représenter par des points du plan. On obtient ainsi une représentation graphique dite « nuage de points ».

Courbe de régression de Y/X Groupons par classes en choisissant des intervalles de classes $X \in [\sigma_{i-1}, \sigma_i]$ étroits mais quand même suffisamment larges (de sorte que, quand une classe n'est pas vide, elle ne soit pas réduite à un ou deux points seulement : on va chercher un point moyen au sein de chaque classe, et la moyenne n'a de sens que lorsqu'on a suffisamment de données...). Notons P_i la i-ème classe :

$$P_i = \{j \in [\![1, n]\!] \mid X_j \in [\sigma_{i-1}, \sigma_i]\}$$

Notons aussi $x_i = \dfrac{\sigma_i + \sigma_{i-1}}{2}$ le centre de chaque intervalle de classe. La *courbe de régression de Y/X* est une courbe d'équation $y = f(x)$ obtenue en joignant les points de coordonnées $(x_i, f(x_i))$ où $f(x_i)$ est défini par :

$$f(x_i) = \frac{1}{\mathrm{card} P_i} \sum_{j \in P_i} Y_j$$

Ce sont les points moyens de chaque classe. En échangeant les variables x et y, on obtiendrait de même la *courbe de régression de X/Y*.

Comment mesurer la corrélation entre X et Y ? Il arrive que X et Y soient deux variables *indépendantes* : le nuage de points doit alors ressembler à des blocs rectangulaires de côtés parallèles aux axes de coordonnées[2]. À l'opposé, Y peut être une fonction de X, ou bien X une fonction de Y. Si $Y = f(X)$, le nuage de points est entièrement contenu dans la courbe d'équation $y = f(x)$ qui est alors la courbe de régression de Y/X. Mais X et Y peuvent fortement dépendre l'une de l'autre sans que cette dépendance soit de type fonctionnel. On va introduire deux quantités mesurant la dépendance ou « corrélation » entre X et Y.

Définition 15.1 *Le* rapport de corrélation de Pearson *est le nombre*

$$\eta_{Y/X} = 1 - \frac{\sum f_i s_i^2}{s_Y^2}$$

où s_Y^2 est la variance de Y, et s_i^2 sa variance intra-classe au sein de la i-ème classe.

Proposition 15.1 *Le rapport de corrélation de Pearson vérifie les propriétés suivantes :*

(i) $0 \leq \eta_{Y/X} \leq 1$

(ii) $\eta_{Y/X} = 0$ *si et seulement si la courbe de régression de Y/X est une droite parallèle à l'axe des abscisses.*

(iii) *si* $\eta_{Y/X} = 1$, *alors y est une fonction de X.*

(iv) *si Y est une fonction de X, alors* $\eta_{Y/X} \simeq 1$.

Démonstration. Rappelons que

$$s_Y^2 = \sum f_i s_i^2 + \sum f_i (f(x_i) - m_Y)^2$$

où m_Y est la moyenne de Y, et $f(x_i)$ sa moyenne au sein de la i-ème classe (et c'est aussi l'ordonnée d'un point de la courbe de régression de Y/X). On a donc bien toujours

$$\sum f_i s_i^2 \leq s_Y^2$$

donc $0 \leq \eta \leq 1$, et les cas extrêmes sont faciles à caractériser.

2. Dans ce cas, les courbes de régression de Y/X et de X/Y sont deux droites parallèles aux axes.

Remarque $\eta_{Y/X}$ et $\eta_{X/Y}$ peuvent être tous deux nuls bien que Y dépende fortement de X sans toutefois être fonction de X : il suffit que le nuage ait deux axes de symétrie parallèles aux axes Ox, Oy.

Définition 15.2 *Le* coefficient de corrélation linéaire *est le nombre*

$$\rho = \frac{\operatorname{cov}(X,Y)}{s_X s_Y}$$

où s_X et s_Y sont les écarts-types des deux variables et leur covariance $\operatorname{cov}(X,Y)$ *est définie par :*

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum (X_i - m_X)(Y_i - m_Y)$$

où m_X et m_Y sont les moyennes des deux variables.

Proposition 15.2 *Le coefficient de corrélation linéaire vérifie les propriétés suivantes :*

(i) $|\rho| \leq 1$

(ii) $|\rho| = 1$ *si et seulement si Y est fonction affine de X, de la forme $Y = aX + b$, avec $a \neq 0$.*

(iii) Si la courbe de régression de Y/X est une droite parallèle à l'axe des abscisses, alors $\rho \simeq 0$

Démonstration. Posons $U = X - m_X$ et $V = Y - m_Y$. Les deux premières propriétés découlent aisément de l'*identité de Lagrange* :

$$\sum U_i^2 \sum V_i^2 = \Big(\sum U_i V_i\Big)^2 + \sum_{i<j} (U_i V_j - U_j V_i)^2$$

À présent, supposons que la courbe de régression de Y/X est une droite parallèle à l'axe des abscisses. Alors $f(x_i) =$ constante, et

$$\begin{aligned}
\sum (X_i - m_X)(Y_i - m_Y) &\simeq \sum_i \left((x_i - m_X) \sum_{j \in P_i} (Y_i - m_Y) \right) \\
&= \sum_i (x_i - m_X)(f(x_i) - m_Y) \\
&= (f(x_i) - m_Y) \sum_i (x_i - m_X) \\
&= 0
\end{aligned}$$

donc $\operatorname{cov}(X,Y) \simeq 0$ et $\rho \simeq 0$.

Remarque On voit que

$$[|\rho| = 1] \Rightarrow \left[\eta_{Y/X} \simeq 1\right],$$

et $\eta_{Y/X} \simeq 1$ entraîne que Y est *presque une fonction de X*. D'autre part :

$$\left[\eta_{Y/X} = 0\right] \Rightarrow [\rho \simeq 0],$$

mais ceci ($\rho \simeq 0$) *ne* permet *pas* de conclure que X et Y sont *presque indépendantes* (penser à une nuage de points ayant deux axes parallèles à Ox, Oy comme mentionné ci-dessus). Pourtant, $\eta_{Y/X} = 0$ permet tout de même de conclure que Y *n'est pas* une fonction de X (sauf si c'est une fonction constante).

Conclusion Le rapport $\eta_{Y/X}$ permet de tester si Y est ou n'est pas une fonction de X. Le coefficient ρ permet de tester si Y est ou n'est pas fonction affine de X. Ils permettent donc parfois tous deux de conclure que Y n'est pas indépendante de X, mais ils ne suffisent jamais à conclure que X et Y sont indépendantes.

15.2 Droite des moindres carrés

On cherche à construire une droite d'équation $y = ax + b$ qui reflète assez bien la forme du nuage de points. À cet effet, on cherche des valeurs de a et b minimisant cette somme de carrés :

$$\sum_{i=1}^{n}(Y_i - aX_i - b)^2$$

Dans un premier temps, on supposera que chacune des deux variables est de moyenne nulle : $m_X = m_Y = 0$. En développant alors la somme des carrés, on trouve :

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - aX_i - b)^2 &= nb^2 + \sum_{i=1}^{n}(Y_i - aX_i)^2 \\
&= nb^2 + a^2\sum X_i^2 - 2a\sum X_iY_i + \sum Y_i^2 \\
&= nb^2 + ns_X^2\left(a^2 - 2a\frac{\operatorname{cov}(X,Y)}{s_X^2} + \frac{s_Y^2}{s_X^2}\right) \\
&= nb^2 + ns_X^2\left(\left(a - \frac{\operatorname{cov}(X,Y)}{s_X^2}\right)^2 + ...\right)
\end{aligned}$$

Cette expression atteint sa valeur minimale quand

$$\begin{cases} a = \dfrac{\operatorname{cov}(X,Y)}{s_X^2} \\ b = 0 \end{cases}$$

Pour deux variables X, Y quelconques, on se ramène au cas précédent en faisant un changement de variables $u = x - m_X, \quad v = y - m_Y$. L'équation de la *droite des moindres carrés* est donc :

$$y - m_Y = \frac{\operatorname{cov}(X,Y)}{s_X^2}(x - m_X)$$

Proposition 15.3 *Si la courbe de régression de Y/X est une droite, alors cette droite est la droite des moindres carrés.*

On appelle parfois la droite des moindres carrés « droite de régression de Y/X ».

Démonstration. La variance intra-classe de Y au sein de la i-ème classe est égale à la variance de $(Y - ax_i - b)$ au sein de la i-ème classe :

$$s_i^2 = \frac{1}{\operatorname{card}P_i}\left(\sum_{j\in P_i}(Y_j - ax_i - b)^2\right) - (f(x_i) - ax_i - b)^2$$

Dans la somme des carrés à minimiser pour avoir la droite des moindres carrés, les termes relevant de la i-ème classe se décomposent donc ainsi :

$$\sum_{j\in P_i}(Y_j - ax_i - b)^2 = \operatorname{card}(P_i) \times \left(s_i^2 + (f(x_i) - ax_i - b)^2\right)$$

et cette expression atteint son minimum quand $f(x_i) = ax_i + b$.

15.3 Exercices

Exercice 15.1 *Voici des données sur un échantillon de 11 individus :*

X	7	5	8	3	6	10	12	4	9	15	18
Y	21	15	24	9	18	30	36	12	27	45	54

Calculer le coefficient de corrélation linéaire. Y a-t-il une forte relation entre X et Y ?

Exercice 15.2 *Le tableau suivant donne l'indice mensuel des dépenses d'assurance maladie d'août 2002 à juin 2003.*

mois	*août 2002*	*octobre 2002*	*décembre 2002*	*février 2003*	*avril 2003*	*juin 2003*
rang du mois x_i	1	3	5	7	9	11
indice y_i	$123,4$	$125,9$	$127,5$	$127,9$	129	$131,4$

Représenter le nuage de points $M_i(x_i, y_i)$ associé à la série statistique. G désigne le point moyen du nuage, c'est le point de coordonnées $(\bar{x}, \bar{y})$. On veut réaliser un ajustement affine de ce nuage de points.

1. *Déterminer les coordonnées du point G et placer ce point sur le graphique.*
2. *Calculer le coefficient de corrélation linéaire ρ et interpréter le résultat obtenu.*
3. *Le modèle étudié dans cette question est appelé « droite de Mayer ». G_1 désigne le point moyen des trois premiers points du nuage et G_2 celui des trois derniers points. Déterminer les coordonnées de G_1 et G_2, puis l'équation $y = Ax + B$ de la droite (G_1G_2). La tracer. Déterminer la somme des résidus pour cet ajustement affine :*

$$S_1 = \sum (y_i - Ax_i - B)^2$$

4. *Le deuxième modèle proposé est celui dit « des moindres carrés ». La « droite de régression de y en x » est alors la droite D d'équation :*

$$y - y_G = m(x - x_G)$$

où le coefficient directeur m est :

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sigma^2(x)}$$

Calculer m et tracer D. Montrer que la somme des résidus, pour cet ajustement, est $S_2 = 1,7$.

5. *Des droites D et (G_1G_2), quelle est celle qui réalise le meilleur ajustement affine ?*
6. *Quels sont les indices mensuels que l'on pouvait prévoir en utilisant l'ajustement affine par la méthode des moindres carrés pour les mois cités dans le tableau suivant ? Commenter les résultats observés.*

mois	*nov 2003*	*déc 2003*	*jan 2004*
indices prévisionnels			
tendances réellement observées	$134,3$	$133,4$	$133,5$

16 Annexe : problème

Le but de ce travail est de commenter mathématiquement les deux articles cités ci-dessous, en répondant aux questions suivantes (il est fortement conseillé de suivre l'ordre des questions). On commencera par lire les deux articles.

1) Qu'est-ce que le *taux de suicide* ?

2) Parmi 100000 actifs, on compte en moyenne 10000 sans emploi dont 5 se suicident par an. Au sein d'une telle population de 100000 actifs, on choisit un individu aléatoirement. On considère les deux événements suivants :

$$A = \text{« l'individu choisi se suicide au cours de l'année »}$$

$$B = \text{« l'individu choisi est sans emploi »}$$

A et B sont-ils des événements indépendants ?

3) On considère un échantillon aléatoire de 100000 individus d'âge actif au sein d'une vaste population (la population française). Notons X le nombre de suicides parmi ces 100000 individus, en une année donnée. Quelle est la loi de probabilité de X ?

4) Au moyen d'un générateur aléatoire, réalisez un algorithme simulant les suicides au sein de cet échantillon de 100000 individus. Chaque suicide devra être réalisé comme une expérience aléatoire de Bernoulli avec deux issues possibles, et une boucle comptera le nombre de suicides à peu près ainsi :

```
for (i=0; i<100000; i=i+1) if (random()>taux) X=X+1;
```

Votre programme devra ensuite répéter cette expérience un grand nombre de fois (cent fois par exemple), et faire des statistiques sur la variable X : il devra tracer l'histogramme de X.

5) Montrez empiriquement que la variable X suit approximativement une loi normale dont vous indiquerez les paramètres. Tracez à cet effet la courbe représentative de la fonction densité d'une telle loi, et comparez-la à l'histogramme de la question précédente.

6) Notons à présent X le nombre de suicides parmi les 100000 salariés de France Télécom, en une année donnée. En vous aidant de la question précédente, et sous l'hypothèse qu'« il n'y a pas de vague de suicide », c'est-à-dire, comme le laissent entendre ces deux articles, sous l'hypothèse que le taux de suicide théorique chez France Télécom est égal au taux de suicide dans la population d'âge actif, calculer un intervalle tel qu'on ait, approximativement,

$$p(X \in [19,6 - ...; 19,6 + ...]) \geq 95\%$$

Peut-on rejeter l'hypothèse ci-dessus sur la base de ce calcul et des statistiques de 2009 ? Peut-on pour autant en conclure qu'« il n'y a pas de vague de suicide » ?

7) À présent, admettons que la loi de Poisson est une meilleure approximation de la loi de X. Quel doit être le paramètre λ de cette loi de Poisson ? Au moyen de la loi de Poisson, tabulez toutes les valeurs de $p(X = k)$ pour k allant de 0 à 30, puis calculez la fonction de répartition de X sur l'intervalle $[0, 30]$, et trouvez le plus petit entier k tel que

$$p(X \leq k) \geq 95\%$$

Enfin, calculez $p(X \in [11, 28])$ et $p(X \in [12, 27])$. Qu'en déduire ?

8) Cherchez et analysez d'autres données et des cas semblables dans la presse (suicides chez Renault en 2007, le suicide dans la Police nationale en 1996,...)

9) Êtes-vous d'accord avec les critiques exprimées dans le deuxième article ? Pouvez-vous formuler vous-même d'autres critiques concernant le premier ou le deuxième article ? La normalité peut-elle être testée par un seuil quantitatif ?

***Sur une vague de suicides*, par René Padieu (La Croix 20/10/2009)**

> "On se suicide plutôt moins à France Télécom qu'ailleurs. Et, semble-il, moins qu'il y a quelques années. Il n'y a pas de 'vague de suicides'", estime René Padieu, inspecteur général honoraire de l'Insee, président de la commission de déontologie de la Société française de statistique.
>
> Depuis quelques semaines, les médias parlent d'une vague de suicides à France Télécom : 24 en dix-neuf mois. Télévision, journaux gratuits ou grands quotidiens nationaux, en passant par d'innombrables forums Internet. On dénombre chaque nouveau cas : le nombre est donc important ! Pourtant, personne ne semble penser à vérifier en quoi il est élevé. Un journaliste consciencieux recoupe son information. Tout citoyen trouve presque tout sur Internet.
>
> En 2007 (cela varie peu d'une année à l'autre), on avait pour la population d'âge actif (20 à 60 ans) un taux de 19,6 suicides pour 100 000. [3] Vingt-quatre suicides en dix-neuf mois, cela fait 15 sur une année. L'entreprise compte à peu près 100 000 employés. Conclusion : on se suicide plutôt moins à France Télécom qu'ailleurs. Et, semble-il, moins qu'il y a quelques années. Il n'y a pas de « vague de suicides »...
>
> Le métier des télécommunications est particulièrement pénible ? La direction de France Télécom est particulièrement inhumaine ? Le stress au travail pousse au suicide ? Possible : mais comme, au final, les gens de France Télécom ne se suicident pas plus que les autres, de ces trois suppositions l'une au moins n'est pas exacte.
>
> "On regrettera que les suicides soient instrumentalisés" Pareillement, en 1996, les suicides dans la police défrayaient la chronique. On expliquait : les policiers font un métier éprouvant et ils ont leur arme de service sous la main. L'association Pénombre avait révélé qu'ils se suicidaient autant, pas plus, pas moins, que leurs compatriotes. [4]
>
> Donc : circulez, il n'y a rien à voir ? Eh bien, si ! Et ceci devient intéressant. Notons d'abord qu'un chiffre, qui n'a rien d'extraordinaire, est brandi pour argumenter un problème : quelle objectivité est-il censé apporter ? Ce qui fait sens ici n'est pas le chiffre lui-même, mais le fait de l'invoquer. Relevons que la révélation des suicides en cause suit la création, par un syndicat, d'un « observatoire du stress » : quand on se met à observer quelque chose, on le voit apparaître. Notons aussi que tous les cas sont imputés à l'entreprise.
>
> Or, le suicide est une rupture (un « passage à l'acte ») dans une tension d'origines multiples. Le stress est un état – pour une cellule, un organisme entier, voire un groupe – où les ressources qui permettent de surmonter un besoin ou une agression séparément viennent à manquer pour en affronter trop à la fois : l'une n'est pas plus la cause qu'une autre. Seule émerge la dernière venue, celle qui, dans la conscience du sujet, a rendu la situation intenable. (Le climat social a même pu contribuer à l'en convaincre : à France Télécom, il est visiblement détestable.)

3. Statistique sur les causes médicales de décès, Inserm, CépiDoc.

4. Nicolas Bourgoin, « Le Suicide dans la police », Pénombre Lettre grise n° 3, printemps 1997. Consultable sur www.penombre.org

En fait, lui ont manqué, en lui ou par ailleurs, les soutiens qui ont permis aux autres d'affronter la même situation. On regrettera que les drames humains que sont ces suicides – peu nombreux, certes, mais bien réels – soient instrumentalisés dans l'affrontement entre une direction et ses salariés : c'est indigne.

"C'est le corps social qui délire" On s'étonnera enfin de voir les jugements sommaires, les explications péremptoires, les propos haineux qui fleurissent sur les plateaux télé et forums Internet. Les responsables industriels ou politiques n'osent pas, ne peuvent pas dire les choses comme elles sont : car la clameur rend sourd et recouvre ce qu'on ne sait examiner sereinement.

Croire en l'existence de quelque chose qui n'est pas constitue ce qu'en psychiatrie on appelle un délire. Ici, ce n'est personne en particulier, mais le corps social qui délire : salariés, direction, ministre, syndicat, journalistes, commentateurs, vous et moi tous ensemble. Ce qui est dit dans ce délire n'est pas réel : c'est quand même un symptôme. Il signe quelque chose, un mal-être social.

Séparément, ergonomes, sociologues et psychiatres dissertent gravement des trois thèmes cités plus haut : mutation des métiers, management, effets du stress. Mais aucune discipline n'analyse – ni donc n'aide à contrôler – cette dynamique politique folle, ce maniement médiatique d'affirmations controuvées et de jugements abrupts à propos d'un problème réel.

***Suicides et statistiques*, Frédéric Lemaître (Le Monde, 24/10/2009)**

Français, on vous ment ! Depuis quelques semaines, on vous fait croire qu'il y a une vague de suicides chez France Télécom. C'est faux ! La direction vient-elle d'envoyer un questionnaire très instructif « sur le stress et les conditions de travail » pour tenter d'enrayer le phénomène (lire sur lemaitre.blog.monde.fr) ? Elle a eu tort ! Dans La Croix, le président de la commission de déontologie de la société française de statistique rétablit la vérité.

Sachant que le taux de suicides des 20-60 ans est de 19,6 pour 100 000 personnes par an, « vingt-quatre suicides en dix-neuf mois, cela fait quinze sur une année. L'entreprise compte près de 100 000 employés. Conclusion : on se suicide plutôt moins à France Télécom qu'ailleurs ». Comment expliquer alors l'émotion suscitée par ce qui se passe dans cette entreprise ? Par un délire collectif. Purement et simplement. « Salariés, direction, ministre, syndicats, journalistes, commentateurs », le « corps social délire », explique cet inspecteur général honoraire de l'Insee.

Et si c'était l'inverse ? Si cette polémique montrait justement la limite de la statistique ? Quantitativement, le taux de suicides n'est pas supérieur à France Télécom que dans le reste de la population d'âge actif. Dont acte. Pourtant, les documents laissés par les personnes qui en viennent à cette extrémité, les conditions dans lesquelles elles ont mis fin à leurs jours ou les témoignages de leur entourage semblent mettre en avant l'importance du contexte professionnel. Autant d'éléments qui, certes, ne font pas une preuve, mais qui sont parlants et échappent à toute statistique.

L'expérience similaire vécue par Renault le prouve, les syndicats sont désormais sensibilisés à la question. Qui peut croire que, si une vingtaine de postiers, d'agents de la SNCF ou d'EDF mettaient fin à leurs jours dans des conditions similaires à ceux de France Télécom, les syndicalistes resteraient sans réaction ?

Si le doute subsiste, c'est là encore en raison des limites de la statistique. Les spécialistes évaluent les phénomènes de sous-déclaration des suicides à 20 %, voire 25 %. Nul ne sait donc s'il y a 10 000 ou 13 000 suicides par an en France (revue Etudes et résultats n° 488 de la direction de la recherche du ministère de la santé).

Autre inconnue majeure : les liens entre suicide et travail. « Aucune étude ne recense les causes imputables au travail. Seule une étude réalisée en Basse-Normandie en 2003, à l'initiative de la Fédération française de santé au travail, laisse poindre le chiffre possible de 300 à 400 suicides annuels imputables au travail », indique l'Anact, l'Agence nationale pour l'amélioration des conditions de travail.

Si l'on ne peut exclure un délire du corps social, la probabilité est non nulle que cette polémique donne raison à Benjamin Disraeli, ce premier ministre britannique qui recensait trois sortes de mensonges : « Le mensonge, le gros mensonge et la statistique. »