Société de Calcul Mathématique SA
Outils d'aide à la décision

Données censurées :

Comment les prendre en compte ?

Manuel théorique et pratique

par Bernard Beauzamy, avec la collaboration de Lucie Le Falher

Août 2013

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel : 01 42 89 10 89. Fax : 01 42 89 10 69. www.scmsa.eu
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 7219Z

Préface

Le présent manuel indique comment prendre en compte des données "statistiquement censurées" : c'est un sujet bien technique, sur lequel il ne devrait pas y avoir polémique. Mais, à l'heure actuelle, tout porte à inquiétude : les données ne dissimulent-elles pas quelque effroyable péril, que l'on cherche à nous cacher ? Il y a ceux qui ont peur de leur ombre, ceux qui voudraient mesurer l'ombre, ceux qui voudraient occulter le soleil et ceux qui voudraient fermer les yeux.

Pour étayer ces craintes modernes, pour leur donner un semblant d'apparence scientifique, il y a des modèles, "amas confus de lois faites souvent au hasard et pour un besoin passager, différentes entre elles de province en province, de ville en ville, et presque toujours contradictoires entre elles dans le même lieu".

Il était donc nécessaire d'en revenir aux lois de la Nature, dont les mathématiques sont la traduction en langue quantitative, patiemment forgée en six mille ans de rude discipline intellectuelle. Dans ce que nous présentons ici, rien n'a moins de trois cent cinquante ans ; toutes les lois que nous utilisons ont, de multiples fois, montré leur aptitude à décrire la réalité. Le troisième chapitre, plus spécifiquement, s'appuie sur la Méthode de Pesée d'Archimède.

Celui qui lira ce manuel pourra, en paix avec lui-même et avec sa conscience, attendre de pied ferme l'Autorité de Sûreté la plus pointilleuse, la Commission Européenne et ses décrets les plus alambiqués, l'ayatollah le plus convaincu des économies d'énergie et du développement durable. Toutes les excommunications lancées par tous ces Torquemada des temps modernes échoueront, comme d'insignifiantes fléchettes, devant l'édifice inébranlable qui est le nôtre.

Mais à l'inverse, à qui ne le lira pas, nous dirons avec Gabriel Marcia Marquez ("Cien años de soledad") : à tous ceux qui sont condamnés à cent ans de solitude, il ne sera pas sur la terre donné de seconde chance. Isolés dans leurs chapelles, ils seront emportés par d'interminables querelles de nature théologique, dont aucun énoncé clair ne sortira jamais.

Le présent manuel est organisé en trois chapitres :

- Le premier présente le sujet ;
- Le second concerne le cas, très fréquent en pratique mais jamais convenablement traité en théorie, où toutes les données sont censurées ;
- Le troisième concerne la situation mixte : on dispose à la fois de données non censurées et de données censurées. Nous présentons la méthode la plus utilisée dans ce cas, surtout dans le milieu médical : méthode dite de "Kaplan-Meier", dont nous montrons qu'elle n'est correcte ni dans son principe ni dans son application : elle relève du bricolage. Nous décrivons explicitement une méthode reposant sur une application appropriée des probabilités conditionnelles.

Chapitre I

Présentation du sujet

I. Définitions préliminaires

Une donnée est dite "censurée" si on n'en connaît pas la valeur exacte, mais seulement une estimation, inférieure ou supérieure, c'est-à-dire une information grossière, du type $X \geq c$ ou $X \leq c$. Une telle information est très pauvre, plus pauvre que de dire "X est entre a et b", puisqu'une seule des deux bornes est connue.

Le mot "censure" est ici d'usage statistique ; il n'a pas grand'chose à voir avec une commission de contrôle, qui aurait décidé de tronquer les données pour que le public n'en ait pas connaissance, encore que, comme nous le verrons, ceci se produise dans certains exemples. Il aurait certainement été préférable de parler de "données tronquées", mais le mot "censure" est consacré par l'usage, et c'est lui que nous emploierons.

On peut ranger les situations où l'on trouve des données censurées en deux classes :

– Celles où les données réelles existent, mais n'ont pas été utilisées ;

– Celles où les données réelles n'existent pas.

Dans nos travaux au quotidien, nous avons rencontré quatre exemples de données censurées ; deux appartiennent à la première catégorie et deux à la seconde. Il en existe d'autres, mais ces exemples suffisent à montrer que le sujet se rencontre souvent.

A. Les données existent

1. Données liées à l'environnement

Il s'agit par exemple de données de radioactivité, dans l'air, dans des fûts ou containers quelconques. On a mesuré l'activité de certains radionucléides, mais lorsque cette activité est inférieure au seuil réglementaire, on ne publie pas la valeur réelle, mais seulement l'information "inférieure au seuil". Cette situation se rencontre aussi dans bon nombre de mesures liées à l'environnement (niveaux de pollution, etc.). La véritable valeur de la mesure est généralement perdue ; plus exactement, elle n'est pas conservée, parce que les responsables n'en voient pas l'intérêt.

2. Données liées aux assurances

Pour pratiquement tous les types de sinistre, les compagnies d'assurance souscrivent une "réassurance" auprès d'une compagnie spécialisée. Cela signifie que si un montant de remboursement dépasse un certain seuil, il est transféré à la compagnie de réassurance, et la compagnie d'assurance elle-même, dans ses bilans, ne fait figurer que la partie inférieure au seuil. Ceci est tout à fait légitime sur le plan comptable, puisque c'est la seule partie qui la concerne ; le vrai coût du sinistre (coût total assurance plus réassurance) est cependant conservé, dans des fichiers séparés.

B. Les données n'existent pas

1. Dans le milieu médical

Il est fréquent de voir un médicament essayé sur un panel (malades ou bien-portants), mais, pour une raison ou une autre, certaines personnes quittent le panel avant la fin de l'expérience et sont perdues de vue : imaginons que 20 000 personnes dans une ville soient suivies pour un traitement contre l'obésité. Si quelqu'un quitte la ville avant la fin de l'observation, les données le concernant ne sont pas complètes. Si l'observation porte sur la survie d'un patient, on pourra dire seulement dans ces conditions que sa durée de vie a été supérieure ou égale à ce qui a été observé.

On pourrait, dans certains cas, se prémunir contre la "fuite" du patient, en lui demandant par avance, où qu'il aille, de bien vouloir donner de ses nouvelles. Mais les conditions de vie, d'alimentation, etc., risquent d'être différentes et ce patient ne sera plus représentatif de ce que l'on cherche à étudier. Par exemple, si l'on étudie l'effet des lignes à haute tension, un habitant qui quitte ce voisinage ne sera plus inclus dans le panel.

2. Dans le milieu industriel

Imaginons un industriel qui fait une expérience quant à la résistance de certaines pièces ; elles sont insérées dans une machine spéciale qui les soumet à une forte pression. Cette expérience doit normalement durer plusieurs jours. Mais il arrive que, même au bout d'une semaine, la pièce n'ait pas été détruite. L'industriel ne souhaite pas poursuivre l'expérience, pour des raisons qui peuvent être diverses : la durée observée lui convient ; il a besoin de la machine pour autre chose, etc. Dans ces conditions, on dira seulement que la durée de vie de la pièce est supérieure ou égale à la durée observée. Il s'agit d'un exemple très semblable à celui du paragraphe précédent, dans son principe.

Mais outre ces situations très bien décrites, on rencontre beaucoup de cas où l'information disponible est de la forme "au moins telle valeur" ou bien "au plus telle valeur". En voici deux exemples :

– Si on cherche à dénombrer les fraudeurs (fraude fiscale, fraude aux documents administratifs, etc.), l'estimation dont on dispose ne tient compte que des fraudeurs démasqués ; les autres sont en nombre inconnu ;

– Si on veut recenser les bénéficiaires potentiels de systèmes d'aide sociale (restaurants du cœur, allocations diverses), on comptera les ayants-droit qui se manifestent effectivement, mais on ne connaît pas le nombre de ceux qui ne se manifestent pas.

Dans la vie de tous les jours, le nombre de situations où l'on ne dispose que d'une estimation inférieure ou supérieure est donc élevé.

II. Est-il souhaitable de censurer les données ?

La question ne se pose, bien sûr, que dans le cas où les données réelles existent. La réponse est simple et claire : il faut toujours utiliser les données réelles et ne jamais les censurer. Il y a deux raisons à cela :

– Scientifiquement, la censure équivaut à une perte d'information, ce qui est toujours malsain ;

– Socialement, les gens se méfient toujours des affirmations "inférieur à un seuil" et se demandent ce que cela cache : est-ce très inférieur au seuil, ou bien juste en dessous ? Qui a fixé la valeur du seuil ? Comment le sait-il ? Etc.

Donner les valeurs réelles, par exemple d'une pollution, permet en outre d'habituer le public à deux notions qu'il connaît mal :

– L'existence d'une incertitude : une mesure n'est jamais absolument précise ;

– Le résultat de la mesure est souvent naturellement variable, d'un jour à l'autre, d'un endroit à l'autre.

Le public est rarement conscient de cette variabilité : que l'on pense aux mesures de CO2, pour lesquelles on se contente de quelques centaines de stations sur la planète, alors que la concentration varie fortement d'un endroit à l'autre, comme la température.

Pourquoi, dès lors, publie-t-on des données censurées alors qu'on pourrait publier des données réelles ? La réponse est simple : il y a des organismes dont c'est le métier que de "traiter" les données avant de les mettre à disposition du public, et ces organismes ont décidé que la censure était le traitement approprié.

Comme ces organismes tiennent leur légitimité précisément de ce traitement, et que leur niveau scientifique d'ensemble est assez faible, on ne peut pas s'attendre à des changements dans l'immédiat, à moins qu'une loi ne vienne dire, comme c'est le cas aux USA,

que les données récoltées avec l'argent public doivent être mises telles quelles, sans aucun traitement, à la disposition du public.

III. Un peu de terminologie

Nous appellerons "donnée réelle" une donnée non censurée ; elle peut bien sûr être sujette à erreur de mesure, incertitude, etc.

Parmi les censures, nous distinguerons entre :

- censure à droite : de la forme $X \geq C$;
- censure à gauche : de la forme $X \leq C$.

Nous ne ferons pas de différence théorique entre inégalité stricte et inégalité au sens large ; dans la pratique, il suffit de déplacer légèrement la borne pour passer de l'un à l'autre.

Pour passer d'une variable censurée à droite à une variable censurée à gauche, ou inversement, il suffit bien sûr de remplacer X par $-X$. Mais ce peut être incommode : si les valeurs mesurées pour X étaient entièrement des nombres positifs (par exemple des durées de vie, des concentrations, etc.), on se retrouve avec des valeurs entièrement négatives, ce qui est déplaisant. Il vaut mieux procéder comme suit.

IV. Conversion d'une censure à droite en une censure à gauche, et réciproquement

- Si B est un nombre qui majore toutes les valeurs censurées à gauche, $X \leq C \leq B$, alors $B - X \geq B - C$ et la variable $Y = B - X$ est positive et censurée à droite ;
- Inversement, si B est un nombre qui majore toutes les valeurs que peut prendre X, $B \geq X$, alors si $X \geq C, B - X \leq B - C$ et la variable $Y = B - X$ est positive et censurée à gauche.

On constate donc que, dans tous les cas, on peut se ramener à une variable positive et censurée à gauche, $X \leq C$, et nous nous limiterons à cette situation dans la suite. Mais attention ! comme nous le verrons par la suite, le choix de B peut n'être pas neutre.

V. Que veut-on faire ?

Il est clair qu'aucune méthode mathématique, si sophistiquée soit-elle, ne peut remplacer des données manquantes. Lorsqu'une information est censurée, la valeur exacte est irrémédiablement perdue. On peut néanmoins, par un travail approprié et en faisant certaines hypothèses, obtenir des résultats de deux types :

– Reconstituer une loi de probabilité pour le phénomène, qui prenne en compte les données censurées ;

– Reconstituer un "tableau d'occurrences", pour le phénomène, qui "ressemble" à celui qu'on aurait eu sans censure.

Nous allons maintenant détailler ces deux aspects, qui sont très différents. Nous verrons que le premier objectif est légitime, et assez facile à réaliser ; le second ne l'est pas, même s'il est souvent demandé. Un nouveau tableau d'occurrences est difficile à réaliser, peut l'être de différentes manières distinctes entre elles, est sujet à caution, et même peut être malhonnête.

VI. Loi de probabilité

On dispose d'une variable à mesurer, X, et on a fait un certain nombre d'observations de cette variable ; à partir de ces observations, censurées ou non, on voudrait se faire une opinion de la loi de probabilité de la variable (voir le livre [MPPR] pour les définitions). Cette loi de probabilité nous renseignera par exemple sur des questions du type : trouver un intervalle de confiance pour la variable à 95%, c'est-à-dire un intervalle tel que les valeurs de X y tombent 95 fois sur 100.

Notre objectif est ici l'obtention d'une loi de probabilité, ce qui ne poserait aucun problème si toutes les données étaient réelles, mais nous avons un certain nombre de données censurées (voire toutes).

Rappelons que la définition d'une loi de probabilité repose sur un découpage des données en classes (appelées "bins" en anglais), de manière à constituer un histogramme (voir [MPPR]). Ce découpage doit être fait en fonction des objectifs, et non en fonction des données. Par exemple, pour une durée de vie, on se demandera : ai-je besoin de l'information heure par heure, jour par jour, ou bien simplement par année ? C'est le besoin qui va conditionner le découpage ; peu importe que les données soient fournies ou non avec vingt chiffres après la virgule.

En d'autres termes encore, le découpage en classes ne peut se faire seulement sur critères mathématiques, mais doit incorporer une analyse du besoin. Ceci est très important et est trop souvent ignoré.

Comme pour nous la valeur des classes n'a aucune importance, nous dirons qu'elles sont représentées par des entiers $1, 2, ..., K$.

Le découpage en classes répond à la question des incertitudes sur les données : toutes les données à l'intérieur d'une même classe sont considérées comme ayant la même valeur (le centre de la classe). En d'autres termes, on "grossit" artificiellement l'incertitude sur les données. Par exemple, on identifie les données correspondant à la même journée, quelle que soit la mesure de la précision horaire.

VII. Tableau d'occurrences

Un tableau d'occurrences est le résultat d'une expérience : à telle date, ou sur telle personne, la variable X a pris telle valeur ; il y a autant de lignes dans le tableau que de répétitions de l'expérience. Un tel tableau, en soi, n'a rien de probabiliste : c'est un compte-rendu d'une expérience. Il n'y a aucune difficulté si toutes les données sont réelles, malheureusement il se peut se trouver que certaines (voire toutes) sont censurées ; on voudrait néanmoins disposer d'un tableau d'occurrences qui reflète l'expérience réalisée, mais sans contenir de censure.

Ce dernier objectif est souvent réclamé dans la pratique, mais il ne peut se concevoir qu'au travers du premier. Nous allons montrer comment réaliser un tableau d'occurrences qui génère la même loi de probabilité, sans données censurées (et au prix de certaines hypothèses), mais ce n'est qu'un tableau fictif. On ne peut pas dire à la 14ème personne, dont le traitement a duré au moins 57 jours : "nous décidons que votre traitement aura duré exactement 132 jours" : ce serait malhonnête. Il doit donc être clair dès maintenant que la reconstitution du tableau sans valeur censurée ne peut avoir de valeur individuelle, mais uniquement une valeur statistique, c'est-à-dire collective et que ce tableau reconstitué n'est qu'une présentation (purement hypothétique) de la loi de probabilité que l'on veut présenter.

Chapitre II

Cas où toutes les données sont censurées

I. Besoin social et attitude scientifique

On étudie dans ce chapitre la situation où on ne dispose que de données censurées et d'aucune donnée réelle. Comme nous l'avons déjà dit, nous nous restreignons à la situation d'une censure $X \leq C$ et nous numérotons les classes possibles $1, 2, ..., K$.

Lorsque la donnée concerne une situation "socialement sensible", comme par exemple une évaluation de pollution, de radioactivité, de durée de vie, etc., la seule attitude acceptable est de se pénaliser : cela revient généralement à déclarer que la valeur réelle est la valeur censurée.

Si par exemple on observe que la radioactivité en provenance d'un fût vérifie $X \leq a$, il est nécessaire de faire comme si c'était la vraie valeur, faute de quoi on pourrait être accusé de dissimulation. De même, si on sait qu'une durée de vie satisfait $X \geq a$, il faut faire comme si cette durée de vie était exactement a. La raison pour laquelle il est impératif de se pénaliser est que l'on ne dispose pas d'arguments permettant de faire autrement ; ce sera le cas au chapitre suivant.

L'attitude scientifique est tout à fait différente. Si nous n'avons pas d'autre information que $X \leq k$, la seule attitude honnête est de considérer que n'importe quelle valeur $1, 2, ..., k$ est possible, avec la même probabilité. Si par exemple nous avons l'information $X \leq 3$, nous considérons que les valeurs possibles sont 1, 2, 3. Il n'y a pas de raison de privilégier telle valeur, satisfaisant $X \leq k$, plutôt que telle autre. En d'autres termes, nous utilisons la loi uniforme sur l'ensemble $1, 2, ..., k$. La loi uniforme est celle qui représente l'absence d'information, comme c'est le cas ici. L'attitude scientifique et l'attitude sociale sont ici très différentes.

Cette approche est très simple à décrire en théorie, mais elle mène à quatre solutions pratiques distinctes, en ce qui concerne les tableaux d'occurrence, ce qui est tout de même beaucoup ! Nous allons les présenter en détail et les critiquer pas à pas. Nous verrons ensuite que la constitution de la loi de probabilité est beaucoup plus simple.

II. Divers tableaux d'occurrence

Voyons comment modifier le tableau d'occurrence pour incorporer les données censurées.

A. Première possibilité

A chaque fois qu'une donnée censurée $X \leq k$ apparaît, on tire au hasard, selon une loi uniforme, un nombre entre 1 et k et on affecte la donnée à la classe ainsi sélectionnée.

Par exemple, si nous avons l'information $X \leq 8$, nous tirons au hasard un nombre entre 1 et 8 ; disons 2. Alors nous remplaçons, pour cette occurrence, $X \leq 8$ par $X = 2$.

Cette méthode est formellement correcte et respecte bien la loi uniforme. En outre, le nombre total d'expériences est conservé. Mais les résultats dépendent beaucoup des tirages qui sont faits selon la loi uniforme, et, dans le cas d'un petit nombre d'expériences, cela peut "faire pencher la balance" dans un sens où dans un autre. Le résultat est en apparence satisfaisant, parce que nous obtenons immédiatement un tableau d'observations sans valeurs censurées et, si quelqu'un nous critique, nous pouvons répondre : c'est le hasard qui l'a voulu.

Voici un exemple avec quatre classes. On a réalisé 100 observations et le nombre de données censurées pour chaque classe est donné en colonne 2. Après quoi, on fait deux simulations distinctes, selon la loi uniforme, pour répartir ces données censurées entre les classes appropriées. Les résultats sont donnés en colonnes 3 et 4.

classe	données censurées	simulation 1	simulation 2
1	25	49	49
2	21	31	30
3	31	18	10
4	23	2	11

Tableau 1 : deux simulations d'insertion de données censurées

Il est très facile de déterminer la loi de probabilité de l'effectif de la classe 4 : nous commençons avec 23 données censurées et chaque donnée a une probabilité $1/4$ d'être affectée à la classe 4. Nous avons donc pour l'effectif final une loi binomiale :

$$P\{X = k\} = \binom{23}{k}\left(\frac{1}{4}\right)^{k}\left(\frac{3}{4}\right)^{23-k}$$

Voici le graphe de cette loi : valeur de k en abscisse, probabilité associée en ordonnée :

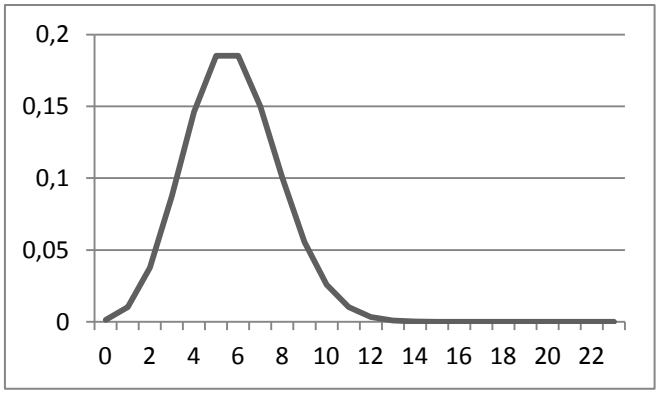

Figure 2 : Graphe de la loi

La loi n'est donc pas très concentrée ; il est normal que l'effectif varie beaucoup d'une simulation à l'autre.

Nous verrons plus loin comment calculer la loi de probabilité des effectifs des autres classes, ce qui est un peu plus difficile.

B. Seconde possibilité

A chaque fois qu'une donnée censurée $X \le k$ apparaît, on répète k fois cette ligne d'observation, une fois avec $X = 1$, une fois avec $X = 2$, etc., une fois avec $X = k$. Si le tableau d'observations initial contenait m_k fois la ligne $X \le k$ $(k = 1,...,K)$ et comportait donc M lignes, $M = m_1 + \cdots + m_K$, alors le tableau final sera beaucoup plus gros : il contiendra M' lignes, avec $M' = m_1 + 2m_2 + \cdots + km_k + \cdots + Km_K$.

Cette méthode n'est pas incorrecte sur le plan théorique, mais elle donne l'impression que le tableau d'observations est beaucoup plus grand qu'il ne l'est en réalité : on a multiplié les lignes de manière factice. Or le nombre d'observations est un élément primordial pour l'évaluation probabiliste, et il ne faut jamais faire semblant d'en avoir plus que l'on en a vraiment (c'est pourquoi les méthodes de type "boostrap" sont sujettes à caution). Nous ne la retiendrons pas, pour cette raison.

Dans le cas de l'exemple donné au tableau 1 ci-dessus, la répétition appropriée de chaque observation conduirait à un tableau de taille $25 + 2 \times 21 + 3 \times 31 + 4 \times 23 = 252$.

C. Troisième possibilité

Elle ressemble en apparence à la seconde. A chaque fois qu'une donnée censurée $X \le k$ apparaît, on répète k fois cette ligne d'observation, une fois avec $X = 1,$ une fois avec $X = 2$, etc., une fois avec $X = k,$ comme précédemment. Mais chacune de ces k lignes va être affectée d'un "poids" égal à $\frac{1}{k}$, ce qui oblige à introduire une colonne supplémentaire appelée "poids de la ligne". Pour faire comprendre ceci sur un exemple concret, imaginons une pièce dont la durée de vie est $X \le 4$. Alors on va considérer que l'on a un quart de pièce dont la durée de vie est 1, un quart dont la durée est 2, un quart dont la durée est 3 et un quart dont la durée est 4. Cela n'a pas de sens physique, mais cela permet d'éviter l'écueil vu précédemment : ici, la taille du tableau d'observations n'augmente plus, mais chacune est découpée en petites fractions. A la fin, on fait la somme pour chaque colonne.

En pratique, bien entendu, on attendra d'avoir lu entièrement le tableau d'observations et de disposer de l'ensemble des nombres $m_1, ..., m_K$. Alors, pour chaque $k = 1, ..., K,$ on répartira, de manière arbitraire, les m_k occurrences de $X \le k$ entre les classes concernées, chacune recevant à peu près $\frac{m_k}{k}$. Si l'on désire des entiers, plutôt que d'arrondir à chaque étape, il vaut mieux le faire à la fin : on aura ainsi moins d'erreurs d'arrondis.

Dans le cas du tableau 1 ci-dessus,

- Les 25 données censurées de la classe 1 seront affectées à la classe 1 ;
- Les 21 données censurées de la classe 2 seront affectées : 10 pour la classe 1, 11 pour la classe 2 ;
- Les 31 données censurées de la classe 3 seront affectées : 10 pour la classe 1, 10 pour la classe 2, 11 pour la classe 3 ;
- Les 23 données censurées de la classe 4 seront affectées : 5 pour la classe 1, 6 pour la classe 2, 6 pour la classe 3, 6 pour la classe 4.

Voici la composition des classes après incorporation :

classe	données censurées	données affectées
1	25	50
2	21	27
3	31	17
4	23	6

Tableau 3 : effectif des classes après incorporation des données censurées

D. Quatrième possibilité

On raisonne de manière intégralement probabiliste ; on considère que l'information $X \le k$ ne doit pas conduire à un tableau déterministe, quel qu'il soit, mais à une loi de probabilité sur l'ensemble des tableaux possibles. En d'autres termes, au lieu d'écrire $X \le k$, on peut écrire l'une quelconque des possibilités $X = 1,...,k$ avec probabilité $\frac{1}{k}$. On fait ceci pour toutes les lignes, et cela nous conduit à un ensemble de tableaux possibles, chacun avec sa probabilité. Nous allons maintenant voir comment réaliser ceci en théorie comme en pratique.

1. Présentation théorique

Commençons par la loi de la classe C_1, qui est la plus complexe de toutes.

On dispose de n_1 données pour la classe C_1 : les données censurées à C_1 sont automatiquement dans cette classe. Les valeurs possibles pour C_1 s'étendent de n_1 à N : nombre total d'observations.

Nous voulons connaître la probabilité que $C_1 = j$, pour $j = n_1,...,N$. Nous allons la calculer par récurrence, en fonction du nombre d'observations faites.

Nous notons $p_n(j)$ la probabilité de la valeur j au terme de la $n-$ème observation ; nous voulons donc calculer $P\{C_1 = j\} = p_N(j)$.

On commence par n_2 observations censurées à C_2. La première d'entre elle peut tomber dans C_1 ou dans C_2 avec probabilité $1/2$. On a donc, au terme de la première observation :

$$p_1(0) = p_1(1) = \frac{1}{2}$$

En effet, au terme de la première observation, la première classe peut être restée vide ou bien contenir un élément.

Pour tout $k > 1$ et tout $j = 0,...,N,$ on a la relation :

$$p_n(j) = a_n\ p_{n-1}(j-1) + (1-a_n)\ p_{n-1}(j) \qquad (1)$$

où a_n est la probabilité que l'élément du $n-$ème tirage tombe dans la classe C_1.

En effet, on peut avoir j éléments à l'étape n de deux manières :

- Ou bien on avait $j-1$ éléments à l'étape précédente, et un élément est arrivé au $n-$ ème tirage ;
- Ou bien on avait déjà j éléments à l'étape précédente, et aucun élément n'est arrivé au $n-$ème tirage ;

Les probabilités "de transition" a_n dépendent de l'étape : pour la classe C_2 elles valent $1/2$, et donc :

$$a_n = \frac{1}{2} \quad \text{pour } n = 1,\ldots,n_2$$

puis elle valent $1/3$ pour C_3 :

$$a_n = \frac{1}{3} \quad \text{pour } n = n_2+1,\ldots,n_2+n_3, \text{ etc.}$$

Le calcul de la famille $p_n(j)$ se fait donc itérativement sur n grâce à la formule (1) et la programmation est facile.

Le même raisonnement s'applique à toutes les autres classes.Voici les lois de probabilité obtenues ainsi pour chacune des classes :

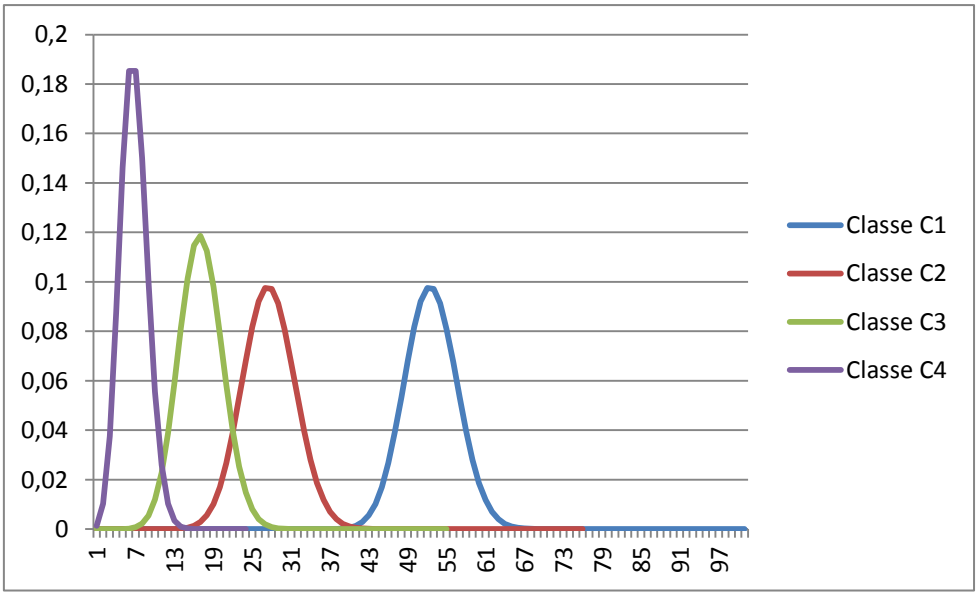

Figure 4 : les lois de probabilité des différentes classes

On constate sur ce graphe que la loi de C_2 est une copie décalée de celle de C_1; ceci est très facile à comprendre : les classes C_4, C_3, C_2 contribuent à C_2 exactement comme à C_1, la seule différence entre les deux est le nombre d'éléments de C_1, qui reste dans cette classe et n'est pas redistribué.

On peut aussi déterminer complètement la loi conjointe de l'ensemble des classes, c'est-à-dire la probabilité :

$$P\{C_1 = j_1, C_2 = j_2, C_3 = j_3\}$$

pour n'importe quel triplet j_1, j_2, j_3 avec $j_1 + j_2 + j_3 = N$. Nous nous limitons ici à trois classes pour simplifier l'écriture.

On introduit en effet :

$$P_n\{C_1 = j_1, C_2 = j_2, C_3 = j_3\}$$

qui est la probabilité à la $n-$ème étape (avec ici $j_1 + j_2 + j_3 = n$). On utilise la relation de récurrence :

$$\begin{aligned} P_n\{C_1 = j_1, C_2 = j_2, C_3 = j_3\} = a_n P_{n-1}\{C_1 = j_1 - 1, C_2 = j_2, C_3 = j_3\} \\ + b_n P_{n-1}\{C_1 = j_1, C_2 = j_2 - 1, C_3 = j_3\} \\ + c_n P_{n-1}\{C_1 = j_1, C_2 = j_2, C_3 = j_3 - 1\} \end{aligned}$$

puisque l'événement $C_1 = j_1, C_2 = j_2, C_3 = j_3$ à l'étape n ne peut provenir que de l'un des trois événements :

$C_1 = j_1 - 1, C_2 = j_2, C_3 = j_3$,
$C_1 = j_1, C_2 = j_2 - 1, C_3 = j_3$,
$C_1 = j_1, C_2 = j_2, C_3 = j_3 - 1$;

Les coefficients a_n, b_n, c_n désignent respectivement la probabilité que la valeur censurée à la $n-$ème observation tombe dans C_1, C_2, C_3.

Bien entendu, pour tout n, $a_n + b_n + c_n = 1$.

2. Implémentation informatique

Dans le cas de la loi d'une seule classe, l'objet de travail, pour la loi de probabilité, est un array :

$$\text{dim } p(1 \text{ to } N) \text{ as double}$$

Dans le cas de la loi conjointe sur trois classes, il nous faut un tableau 2x2, puisque la troisième classe se déduit des deux premières :

$$\text{dim } p(1 \text{ to } N, 1 \text{ to } N) \text{ as double}$$

et on mettra dans $p(i, j)$ la probabilité que $C_1 = i$ et $C_2 = j$.

E. En conclusion

Parmi ces quatre possibilités de traitement, laquelle retenir ? La quatrième est la plus correcte, puisqu'elle aboutit à un ensemble de tableaux, chacun avec sa probabilité. Mais elle est difficile à mettre en œuvre. Celui qui veut un seul tableau, déterministe, incorporant les données censurées choisira la troisième possibilité, malgré son absence de sens physique apparent.

Bien entendu, si le nombre d'observations est grand, ces difficultés disparaissent. Par exemple si $X \leq 4$ a été observé 10 000 fois, il est légitime de dire que chacune des valeurs 1, 2, 3, 4 a été observée 2 500 fois, mais il y a danger à les désigner nominalement.

Lorsqu'on utilise une loi uniforme, comme c'est le cas ici, il faut prendre garde aux bornes qui sont retenues, car ce sont elles qui conditionnent l'aspect de la loi. Ici, nous avons pris des classes de valeur entières, 1, 2, etc., ce qui simplifie la présentation, et si $X \leq 4$, il n'y a que 4 possibilités avant. Mais si la borne est $X \leq 2.356$, le choix de la loi dépendra d'une part de la discrétisation faite (choix de l'histogramme), comme expliqué plus haut, mais aussi de la borne inférieure retenue.

Doit-on considérer que $X \geq 1$, que $X \geq 0$, que $X \geq -50$? Les résultats seront différents. Le choix de la borne inférieure doit être justifié par des considérations liées à la nature des données et, éventuellement, doit faire l'objet d'une analyse de robustesse (on essaye plusieurs choix et on note les différences éventuelles).

III. Loi de probabilité

On peut ne pas vouloir constituer un tableau d'occurrence, mais se contenter d'une loi de probabilité, incorporant les données censurées. C'est à la fois plus facile et plus correct.

On note T la variable de censure (T comme "troncature"). La condition $T = k$ est équivalente à $X \leq k$, mais avec cette notation les événements $\{T = k\}$ sont disjoints : c'est l'indication donnée par la variable de censure.

On a, pour tout j :

$$P\{X = j\} = \sum_{k \geq j} P\{X = j \,|\, Y = k\} P\{Y = k\}$$

puisque $P\{X = j \cap Y = k\} = P\{X = j\}$ si $k \geq j$ et 0 si $k < j$.

Or $P\{X = j \,|\, Y = k\} = \frac{1}{k}$ si $k \geq j$ et 0 si $k < j$, puisque nous avons fait l'hypothèse d'une loi uniforme sur X lorsque la censure est connue.

Par ailleurs, puisque nous avons K classes et que la valeur $T = k$ a été observée m_k fois, on a :

$$P\{T = k\} = \frac{m_k + 1}{M + K}$$

où $M = m_1 + \cdots + m_K$; voir [NMP].Il en résulte que :

$$P\{X = j\} = \sum_{k \geq j} \frac{1}{k} \frac{m_k + 1}{M + K}$$

ce qui nous donne une formule explicite pour la loi de X, après incorporation des données censurées.

IV. Loi conjointe et données censurées

Il peut arriver que deux variables (X,Y) soient observées simultanément, et que chaque observation donne lieu à donnée censurée : $X \le a$ et $Y \le b$. Alors le nombre $m_{a,b}$ d'occurrences de l'observation $X \le a$ et $Y \le b$ doit être réparti de manière aussi uniforme que possible entre les $a \times b$ classes $X \le a$ et $Y \le b$. Par exemple, si $X \le 3$ et $Y \le 2$ a été observé 7 fois, on aura la disposition suivante :

X\Y	1	2
1	7/6	7/6
2	7/6	7/6
3	7/6	7/6

La théorie est la même que précédemment.

Chapitre III

Données réelles et données censurées

I. Besoin social et attitude scientifique

Dans ce chapitre, nous étudions la seconde situation qui se rencontre en pratique : notre tableau d'observations contient à la fois des données réelles et des données censurées. Ici, à la différence du chapitre précédent, nous disposons d'éléments qui nous permettent d'argumenter vis-à-vis du besoin social. Nous avons des données réelles, que nous pouvons présenter, et la question est d'incorporer les données censurées.

Encore faut-il que les données censurées soient obtenues aléatoirement et que la censure soit indépendante du processus qui a permis d'obtenir les données. Si un crétin quelconque a coupé l'électricité un vendredi soir alors qu'on étudiait les durées de vie de certaines pièces, on peut légitimement dire que ces données sont du même genre que celles obtenues un autre jour de la semaine. En revanche, si toutes les données provenant d'une même usine sont censurées, elles ne peuvent être incorporées aux données réelles provenant d'autres usines. On peut dire que, pour cette usine-là, toutes les données sont censurées, et on se retrouve dans la situation du précédent chapitre.

Admettons dans la suite que la censure soit indépendante du processus, que les données réelles existent, et commençons par présenter la méthode la plus utilisée dans ces circonstances, du moins dans le monde médical. Nous en verrons les avantages et les inconvénients, et nous présenterons ensuite les méthodes que nous préconisons.

II. Méthode de Kaplan Meier

Elle a été introduite, dans le milieu médical, pour des durées de vie, qui sont de la forme $X \geq c$ ou $X > c$. Ces durées de vie pouvant prendre une valeur quelconque, une phrase du type "l'évènement n'a pas encore eu lieu à l'instant t équivaut à dire qu'il n'a pas eu lieu juste avant t et n'a pas lieu en t", qui est à la base de la méthode, comme on va le voir, paraît acceptable. Mais, une fois les classes définies, le "juste avant" pose problème. En fait, cette méthode, quoique largement utilisée dans le milieu médical, n'est pas correcte. Mais commençons par la présenter.

Dans ce manuel, nous avons développé la théorie pour des données sous la forme $X \le c$ et c'est ainsi que nous allons poursuivre la présentation. Le passage de l'un à l'autre est facile : il suffit de remplacer X par $B - X$, où B est la plus grande valeur que peut prendre la variable X.

A. Présentation théorique

Comme précédemment, nous avons K classes, numérotées de 1 à K ; la $k-$ ème classe est l'ensemble $k-1 < x \le k.$ Pour chaque $k = 1,...,K,$ nous notons m_k le nombre de données censurées et n_k le nombre de données exactes dans la $k-$ ème classe. Nous notons $M = m_1 + \cdots + m_K$ et $N = n_1 + \cdots + n_K$; ce sont, respectivement, le nombre total de données censurées et le nombre total de données exactes. Nous notons aussi $M_k = m_1 + \cdots + m_k$ et $N_k = n_1 + \cdots + n_k$, $k = 1,...,K.$

Commençons, pour présenter la méthode de Kaplan-Meier, par supposer que toutes les classes contiennent des données exactes (peu importe en quel nombre). Alors on écrit :

$$P(X \le j) = P(X \le j \mid X \le j+1) \times P(X \le j+1) \tag{1}$$

et en réitérant :

$$P(X \le j) = \prod_{k=j}^{K-1} P_k \tag{2}$$

où l'on note :

$$P_k = P(X \le k \mid X \le k+1),\ k = 1,...,K-1.$$

Pour la dernière, $P_{K-1} = P\left(X \le K-1 \mid X \le K\right),$ le conditionnement est automatique, puisque la condition $X \le K$ est toujours satisfaite. Les formules (1) et (2) utilisent simplement la définition des probabilités conditionnelles.

Nous allons maintenant évaluer P_k et c'est là que se situe la difficulté.

Le nombre total de données satisfaisant $X \le k+1$ est, par définition, $M_{k+1} + N_{k+1}$; le nombre de données satisfaisant $X \le k$ est estimé par $M_{k+1} + N_k$: cela revient à dire que toutes les données censurées vérifiant $X \le k+1$ doivent automatiquement vérifier $X \le k$ (ou, en d'autres termes, qu'il n'y en a pas entre k et $k+1$) ; cela peut sembler correct pour des durées de vie, si les classes sont à la seconde près (très bref intervalle de temps), mais c'est évidemment faux en général.

Avec cette approche, on obtient l'estimation :

$$P_k \approx \frac{M_{k+1} + N_k}{M_{k+1} + N_{k+1}} \tag{3}$$

d'où il résulte par (2) :

$$P(X \le j) \approx \prod_{k=j}^{K-1} \frac{M_{k+1} + N_k}{M_{k+1} + N_{k+1}} \tag{4}$$

et par différence :

$$\begin{aligned} P\left(X \in C_j\right) &= P\left(X \le j\right) - P\left(X \le j-1\right) \\ &= \prod_{k=j}^{K-1} \frac{M_{k+1} + N_k}{M_{k+1} + N_{k+1}} - \prod_{k=j-1}^{K-1} \frac{M_{k+1} + N_k}{M_{k+1} + N_{k+1}} \\ &= \left(1 - \frac{M_j + N_{j-1}}{M_j + N_j}\right) \prod_{k=j}^{K-1} \frac{M_{k+1} + N_k}{M_{k+1} + N_{k+1}} \\ &= \frac{n_j}{M_j + N_j} \prod_{k=j}^{K-1} \frac{m_{k+1} + n_k}{m_{k+1} + n_{k+1}} \end{aligned}$$

et le nombre de données par classe, après répartition selon cette méthode, est :

$$e_k = (N + M) \times P\left(X = k\right)$$

Si une classe ne contient pas de données exactes, la méthode ne s'applique pas directement, mais une modification assez simple est faite : on réunit cette classe à une autre, contenant des données exactes. Par exemple, si C_2 ne contient pas de données exactes et C_3 en contient, on considère une seule classe, $C_2 \cup C_3$; la réunion peut porter sur un nombre plus élevé, si les classes contenant des données exactes sont plus rares. Ceci ajoute une erreur supplémentaire à la méthode.

Les classes ne contenant pas initialement de données exactes ne contiennent aucune donnée à la fin. Les effectifs attribués aux classes ne sont pas des entiers ; on peut les arrondir pour leur donner un sens physique.

Nous allons maintenant traiter un exemple.

B. Exemple

On relève 70 observations réelles ou censurées à gauche, comme ci-dessous. Les données censurées ($X \leq a$) sont notées a*. Les voici, mises dans l'ordre croissant :

0.09, 0.11, 0.13, 0.15*, 1.32, 1.33, 1.50*,1.70,1.65,1.77*, 1.78*, 1.81, 2.34*,2.55*, 2.59*, 2.59, 2.63*, 2.87, 2.96*, 3.01, 3.07*,3.15, 3.19, 3.23, 3.27, 3.41, 3.50*, 3.60, 3.81*, 3.96, 4.05, 4.15*,4.22* 4.34, 4.55, 4.60*,4.77, 4.69*, 4.72*, 4.80, 4.82*, 5.10*, 5.25, 5.30, 5.44, 5.50, 5.50, 5.51, 5.66, 6.02, 6.03, 7.10*, 7.15,7.42, 7.44*, 7.50*, 7.62*, 8.03, 8.10*, 8.15, 8.20*,8.96, 9.03, 9.16, 9.27, 9.39*, 9.47, 9.72, 9.96*, 9.97*.

On choisit de ranger ces données dans 10 classes, d'amplitude 1. Le tableau ci-dessous récapitule le nombre de données selon la classe. Par exemple, 0.09 est compté dans la classe 1 car $0 < 0.09 \leq 1$.

classe	nb de valeurs réelles	nb de valeurs censurées X<=j
1	3	1
2	5	3
3	2	5
4	8	3
5	5	6
6	7	1
7	2	0
8	2	4
9	3	2
10	5	3

Tableau 1 : Données réelles et censurées

Voici le tableau détaillant les calculs intermédiaires :

classe	Nk	Mk	Pk	P(x<=j)	P(X=j)	e_j
1	3	1	0,58	0,24	0,24	16,52
2	8	4	0,89	0,40	0,17	11,80
3	10	9	0,73	0,45	0,05	3,33
4	18	12	0,88	0,62	0,16	11,51
5	23	18	0,86	0,70	0,09	5,99
6	30	19	0,96	0,82	0,12	8,19
7	32	19	0,96	0,85	0,03	2,34
8	34	23	0,95	0,88	0,03	2,17
9	37	25	0,93	0,93	0,04	3,15
10	42	28		1,00	0,07	5,00

Tableau 2 : Calculs intermédiaires de la méthode de Kaplan-Meier

Finalement, la nouvelle répartition des données est la suivante :

classe	nb de valeurs réelles
1	17
2	12
3	3
4	12
5	6
6	8
7	2
8	2
9	3
10	5

Tableau 3 : Répartition des données d'après la méthode de Kaplan-Meier

Nous comparerons plus bas ces résultats avec ceux obtenus par une méthode, plus correcte sur le plan théorique, que nous présentons maintenant.

III. Approche en termes de probabilités conditionnelles

L'approche que nous présentons maintenant se définit simplement : les données réelles nous ont permis d'établir une loi de probabilité du phénomène (que cette loi soit grossière ou précise, peu importe), et nous voulons maintenant incorporer les données censurées en tenant compte de cette loi. En d'autres termes, l'introduction des données censurées ne doit pas "abîmer" la loi établie à partir des données exactes.

L'approche retenue s'inspire de la méthode de "pesée" d'Archimède (voir [AMW]) : on dispose d'une loi connue (celle constituée à partir des données exactes) et on veut construire une nouvelle loi (en incorporant les données censurées) de manière que, d'une certaine façon, elles aient le même "poids".

A. La situation de base

Comme précédemment, nous avons K classes, dont les valeurs seront prises égales à $1,2,...,K$ pour simplifier. Les valeurs ne nous intéressent pas : seules les probabilités nous intéressent.

Pour chaque $k=1,...,K$, nous avons n_k valeurs réelles (non censurées), c'est-à-dire pour lesquelles l'observation a donné la valeur $X=k$ et nous avons m_k valeurs censurées, c'est-à-dire pour lesquelles $X \le k$.

Soit $N = n_1 + \cdots + n_K$: nombre total de valeurs réelles et soit $M = m_1 + \cdots + m_K$ le nombre total de valeurs censurées.

Prenons un exemple, afin d'illustrer la théorie. Nous avons 10 classes, avec des valeurs de 1 à 10. Nous avons répété N fois l'expérience, et nous avons obtenu des valeurs réelles $n_1, ..., n_{10}$. Nous obtenons maintenant une valeur, dont nous savons seulement qu'elle est ≤ 3. Que devons-nous en faire ?

La réponse est très simple. A partir de l'information provenant des valeurs réelles, la nouvelle observation ≤ 3 augmente la probabilité d'avoir $X \leq 3$ (puisque nous avons une observation de plus avec cette propriété) et de ce fait diminue la probabilité d'avoir $X > 3.$ Mais cette nouvelle information ne dit rien sur la plage $1,2,3$. Donc, l'information reçue ne doit pas modifier la loi de probabilité conditionnelle sachant $X \leq 3$.

A partir de cette remarque, nous allons montrer comment introduire les données censurées, c'est-à-dire déterminer dans quelle classe elles doivent être rangées.

B. Présentation théorique

Pour commencer, nous avons m_1 données avec $X \leq 1$: ces données doivent obligatoirement être rangées dans la première classe, puisqu'il n'y en a aucune au-dessous.

Considérons maintenant les m_k données vérifiant $X \leq k$; nous voulons les ranger dans l'une des classes $C_1, ..., C_k$. Nous allons voir comment réaliser ceci.

Soit $\nu_{j,k}$ le nombre (inconnu) de données que nous voulons mettre dans la $j-$ème classe, $j = 1, ..., k.$ Bien entendu, $\nu_{1,k} + \cdots + \nu_{k,k} = m_k$.

Nous recherchons d'abord la loi conditionnelle $X \leq k.$ Alors X peut prendre seulement k valeurs, à savoir $1,2,...,k,$ et le nombre d'occurrences est $n_1, ..., n_k$ respectivement. Posons, pour simplifier les notations, $N_k = n_1 + \cdots + n_k$.

Avant d'introduire une quelconque donnée censurée, la loi conditionnelle de la $j-$ème classe, sachant $X \leq k$, c'est-à-dire $P\{X = j \mid X \leq k\}$ est simplement :

$$p_{j,k} = \frac{n_j + 1}{N_k + k} \tag{1}$$

Ceci résulte simplement de la théorie générale présentée dans [NMP], chapitre II, page 34 : si on a k classes, avec n_j valeurs dans chacune, la probabilité de la $j-$ème est :

$$p_{j,k} = \frac{n_j + 1}{\sum_{l=1}^{k} n_l + k}$$

Si maintenant nous introduisons $\nu_{j,k}$ valeurs dans la $j-$ ème classe, cette probabilité devient :

$$q_{j,k} = \frac{n_j + \nu_{j,k} + 1}{N_k + m_k + k}$$

La théorie présentée plus haut dit que, pour tout j, $q_{j,k} = p_{j,k}$, c'est-à-dire :

$$\frac{n_j + \nu_{j,k} + 1}{N_k + m_k + k} = \frac{n_j + 1}{N_k + k}$$

Ceci donne :

$$\nu_{j,k} = \frac{n_j + 1}{N_k + k}\left(N_k + m_k + k\right) - \left(n_j + 1\right)$$

et finalement :

$$\nu_{j,k} = \frac{m_k}{N_k + k}\left(n_j + 1\right) \tag{2}$$

La réponse finale est donc assez simple : les nouvelles valeurs doivent être distribuées proportionnellement aux anciennes, plus une unité. Puisque la loi de probabilité conditionnelle est préservée à chaque étape, on peut commencer avec $k = 1$, ou avec $k = K$, ou dans n'importe quel ordre ; l'ordre des opérations n'a pas d'importance.

A la fin, la $j-$ème classe reçoit un nombre d'éléments égal à ce qu'elle avait au départ (données réelles) plus toutes les incorporations pour $k \geq j$, c'est-à-dire :

$$n_{j,\,final} = n_j + \left(n_j + 1\right)\sum_{k=j}^{K} \frac{m_k}{N_k + k} \tag{3}$$

Remarque

Il n'est pas possible d'introduire les données censurées au fur et à mesure qu'elles apparaissent dans le tableau d'observations. Il faut lire tout le tableau, mettre d'un côté les données réelles, d'un autre côté les données censurées, constituer les diverses probabilités conditionnelles et, à partir de ces lois, introduire les données censurées comme expli-

qué ci-dessus. Autrement dit, le tableau d'observations n'intervient qu'au travers des nombres n_k, m_k ; l'ordre des observations n'intervient pas. On peut par exemple supposer que les données censurées sont les dernières.

C. Un exemple

Reprenons l'exemple avec 10 classes, vu pour la présentation de la méthode de Kaplan-Meier.

classe	nb de valeurs réelles	nb de valeurs censurées
1	3	1
2	5	3
3	2	5
4	8	3
5	5	6
6	7	1
7	2	0
8	2	4
9	3	2
10	5	3

Tableau 4 : données réelles et données censurées

Voici le nombre d'occurrences des valeurs réelles :

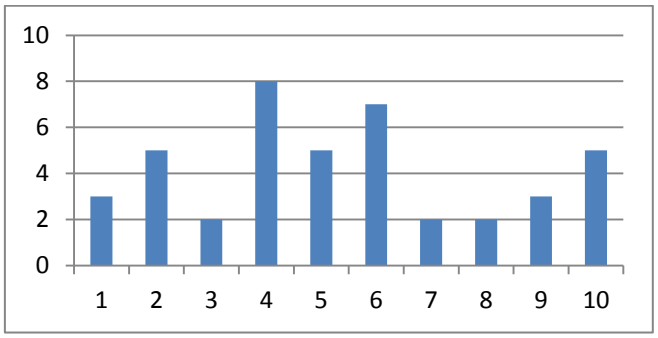

Figure 5 : nombre d'occurrences des valeurs réelles

et voici la loi de probabilité construite sur les valeurs réelles :

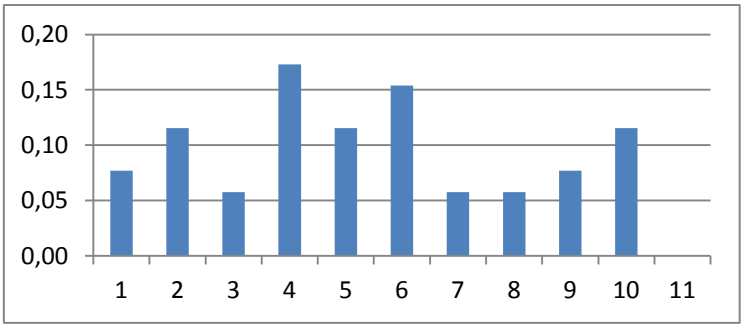

Figure 6 : loi de probabilité des valeurs réelles

Nous allons maintenant incorporer les valeurs censurées. La première valeur, $X \le 1$, doit être mise dans la première classe. La composition des classes devient alors :

classe	nb de valeurs
1	4
2	5
3	2
4	8
5	5
6	7
7	2
8	2
9	3
10	5

Tableau 7 : composition des classes après la première incorporation

Maintenant, nous devons incorporer les 3 valeurs censurées satisfaisant $X \le 2$ et elles doivent être mises dans la première ou dans la seconde classe.

Le nombre à ranger dans la première classe est :

$$\nu_{1,2} = \frac{m_2}{N_2 + 2}(n_1 + 1) = \frac{3}{9+2}(4+1) = 1.36$$

et dans la seconde :

$$\nu_{2,2} = \frac{m_2}{N_2 + 2}(n_2 + 1) = \frac{3}{9+2}(5+1) = 1.64$$

et la somme fait bien 3. Si nous voulons des entiers, nous devrons arrondir et prendre :

$\nu_{1,2} = 1,\ \nu_{2,2} = 2$.

Voici la répartition à chaque étape :

j\k	1	2	3	4	5	6	7	8	9	10	total
1	1,00	1,36	1,76	0,77	1,35	0,19	0,00	0,66	0,32	0,43	7,84
2		1,64	2,35	0,97	1,65	0,24	0,00	0,92	0,43	0,62	8,82
3			0,88	0,39	0,6	0,09	0,00	0,33	0,14	0,19	2,63
4				0,87	1,5	0,20	0,00	0,72	0,35	0,51	4,15
5					0,9	0,13	0,00	0,46	0,20	0,27	1,96
6						0,15	0,00	0,52	0,26	0,35	1,28
7							0,00	0,20	0,09	0,12	0,40
8								0,20	0,09	0,12	0,40
9									0,12	0,16	0,27
10										0,23	0,23
total	1	3	5	3	6	1	0	4	2	3	28

Tableau 8 : répartition des valeurs censurées à chaque étape

Voici la loi de probabilité finale, après incorporation des valeurs censurées :

classe	nb valeurs réelles	nb valeurs ajoutées	total	proba
1	3	7,84	10,84	0,15
2	5	8,82	13,82	0,19
3	2	2,63	4,63	0,07
4	8	4,15	12,15	0,16
5	5	1,96	6,96	0,10
6	7	1,28	8,28	0,12
7	2	0,40	2,40	0,04
8	2	0,40	2,40	0,04
9	3	0,27	3,27	0,05
10	5	0,23	5,23	0,08

Tableau 9 : loi de probabilité finale après incorporation

Voici la comparaison entre la loi d'origine, construite sur les données réelles, et la loi finale, après incorporation des données censurées :

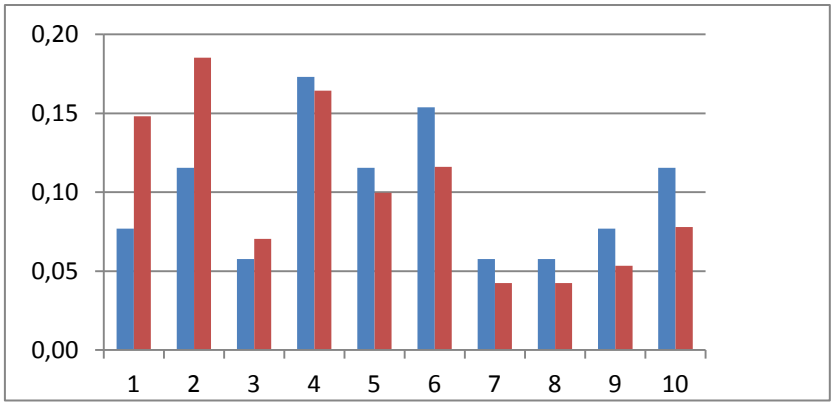

Figure 10 : loi avant et après incorporation des données censurées

En bleu : loi sur données réelles ; en rouge : loi après incorporation des données censurées.

La loi finale peut être très différente de la loi d'origine : seules les lois conditionnelles sont préservées à chaque étape. Par exemple, si nous avons une large quantité de données censurées qui doivent tomber dans l'une des deux premières classes, ces deux classes auront une probabilité beaucoup plus grande à la fin.

Les résultats issus de cette méthode sont différents de ceux obtenus par la méthode de Kaplan Meier. On rappelle ci-dessous les répartitions trouvées :

classe	Kaplan Meier	méthode SCM
1	16,52	10,84
2	11,8	13,82
3	3,33	4,63
4	11,51	12,15
5	5,99	6,96
6	8,19	8,28
7	2,34	2,4
8	2,17	2,4
9	3,15	3,27
10	5	5,23

Tableau 11 : Comparaison de la répartition des données après incorporation des censures selon la méthode utilisée

Nous constatons que les classes supérieures (5-10) présentent pratiquement les mêmes effectifs. En revanche, la répartition varie plus fortement dans les classes inférieures (1-4). Rappelons aussi que si une classe ne contient pas de données exactes, elle n'en recevra pas par la méthode de Kaplan-Meier, ce qui constitue un défaut supplémentaire de cette méthode, dont le principe de base n'est pas correct.

IV. Loi conjointe

Il peut arriver (mais ceci est peu fréquent en pratique) que deux variables ou davantage aient été enregistrées en même temps, et que les données censurées les concernent toutes les deux. Par exemple, nous enregistrons A, B et nous rencontrons une situation où nous savons seulement que $A \le a, B \le b$.

A. Présentation théorique

1. Cas simple

Si nous connaissons une valeur exacte pour B, c'est à dire $A \le a, B = b$, nous appliquons le paragraphe précédent à la loi conditionnelle de A, B sachant $B = b$, c'est-à-dire à une colonne dans la table ci-dessous :

A\B	b1	b2	b3
a1			
a2		X	
a3			

Dans cette table, les X termes avec $A \le a_2, B = b_2$ seront redistribués entre les deux cellules $A = a_1, B = b_2$ et $A = a_2, B = b_2$, c'est-à-dire en ne se servant que de la seconde colonne.

2. Cas général

Nous allons donc traiter le cas où l'inégalité concerne à la fois A et B :

A\B	b1	b2	b3
a1			
a2			
a3		X	

Dans cette situation, les X termes avec $A \le a_3$ et $B \le b_2$ doivent être redistribués entre les six cellules à gauche et au dessus.

La théorie est exactement la même que dans le paragraphe précédent : la loi conditionnelle, sachant $A \le a_3$ et $B \le b_2$ doit être conservée.

Pour présenter le raisonnement correctement, fixons nos notations.

Dans le cas général, nous avons L variables X_l, $l = 1,...,L$ et nous observons leur loi conjointe. Chaque classe est caractérisée par une valeur $b_{k_1,...,k_L}$, où $k_1 = 1,...,K_1$, …, $k_L = 1,...,K_L$ (le centre de la classe). Nous sommes dans un espace de dimension L et l'histogramme pour la loi conjointe est fait de $K_1 \times \cdots \times K_L$ cellules (le nombre de classes n'a pas de raison d'être le même pour toutes les variables).

Soit $n_{k_1,...,k_L}$ le nombre de valeurs réelles (non censurées) appartenant à la classe $b_{k_1,...,k_L}$ et soit $m_{k_1,...,k_L}$ le nombre de valeurs censurées correspondantes ; on doit les distribuer entre toutes les classes $b_{j_1,...,j_L}$ avec $j_1 \le k_1,..., j_L \le k_L$.

Dans une classe $b_{k_1,...,k_L}$, nous avons $m_{k_1,...,k_L}$ données censurées à distribuer entre les classes qui la précèdent. Soit $\nu_{j_1,..,j_L}$ le nombre de données que nous voulons mettre dans la classe $b_{j_1,...,j_L}$ (ce nombre est inconnu).

Nous nous restreignons au "parallélipipède" $b_{j_1,...,j_L}$ avec $j_1 \le k_1,..., j_L \le k_L$.

Introduisons $N_{k_1,..,k_L} = \sum_{j_1 \le k_1,..., j_L \le k_L} n_{j_1,...,j_L}$: c'est le nombre total d'éléments dans le parallélépipède (cette notation est semblable à celle de N_k au paragraphe précédent).

Avant d'avoir incorporé les données censurées, la probabilité de la classe $b_{j_1,...,j_L}$ est :

$$p(j_1,..., j_L) = \frac{n_{j_1,...,j_L} + 1}{N_{k_1,...,k_L} + k_1 \times \cdots \times k_L}$$

En effet, c'est simplement la formule (1) ci-dessus, puisque :

- $n_{j_1,...,j_L}$ est le nombre de termes dans la classe ;
- $N_{k_1,...,k_L} = \sum_{j_1 \le k_1,..., j_L \le k_L} n_{j_1,...,j_L}$ est le nombre total de termes dans le parallélépipède ;
- $k_1 \times \cdots \times k_L$ est le nombre total de classes dans le parallélépipède.

Si nous mettons $\nu_{j_1,..,j_L}$ données censurées dans chaque classe $b_{j_1,...,j_L}$, la probabilité de chaque classe du parallélépipède devient :

$$q(j_1,...,j_L)=\frac{n_{j_1,...,j_L}+\nu_{j_1,..,j_L}+1}{N_{k_1,...,k_L}+m_{k_1,...,k_L}+k_1\times\cdots\times k_L}$$

et les deux expressions doivent être égales, ce qui donne la relation :

$$\frac{n_{j_1,...,j_L}+\nu_{j_1,..,j_L}+1}{N_{k_1,...,k_L}+m_{k_1,...,k_L}+k_1\times\cdots\times k_L}=\frac{n_{j_1,...,j_L}+1}{N_{k_1,...,k_L}+k_1\times\cdots\times k_L}$$

d'où nous déduisons :

$$\nu_{j_1,..,j_L}=\left(n_{j_1,...,j_L}+1\right)\frac{m_{k_1,...,k_L}}{N_{k_1,...,k_L}+k_1\times\cdots\times k_L}$$

ce qui est complètement identique, avec des notations appropriées, à ce que nous avions obtenu au paragraphe précédent.

B. Mise en œuvre pratique

Supposons que nous ayons L variables X_l, $l=1,...,L$; disons $L=6$. Supposons aussi pour simplifier que chacune puisse prendre $K=10$ valeurs. Alors, l'histogramme pour la loin conjointe est défini par K^L cellules, et dans chaque cellule $\left(b_{j_1},...,b_{j_L}\right)$, nous mettons le nombre de fois où l'information $X_1=b_{j_1},...,X_L=b_{j_L}$ a été observée : c'est la définition ordinaire d'un histogramme. Donc, cet histogramme est constitué en balayant la table d'observations et en augmentant à chaque fois la cellule correspondante.

Lorsque nous avons des données censurées, une ligne du tableau d'observations peut avoir la forme suivante :

$$X_1=ou\le b_{j_1},\ X_2=ou\le b_2,...,\ X_L=ou\le b_{j_L}$$

Toutes les comparaisons rencontrées peuvent être $=$ ou $\le$.

Pour chaque j où nous rencontrons le symbole $\le$, nous disons que l'observation a été censurée pour le $j-$ème axe ; nous n'avons que ces axes à prendre en compte.

Pour tout axe censuré, nous construisons le parallélépipède correspondant. Voyons sur un exemple comment procéder. Supposons que nous ayons $K=10$ classes.

Supposons par exemple que le symbole $\le$ apparaisse seulement aux rangs 2, 4, 6, et que notre observation ressemble à ceci :

$$X_1 = 2, X_2 \le 4, X_3 = 1, X_4 \le 3, X_5 = 6, X_6 \le 2$$

Alors les axes censurés sont les numéros 2,4,6. Les données censurées doivent être redistribuées sur un parallélépipède, fait de la façon suivante :

$$X_2 = 1,2,3,4\,,\ X_4 = 1,2,3\,,\ X_6 = 1,2\,.$$

Notre parallélépipède contient donc $4 \times 3 \times 2 = 24$ cellules. La manière de répartir les données entre ces cellules est décrite dans le paragraphe précédent. Cela peut se faire observation par observation (mais alors on obtient des valeurs fractionnaires), soit, de préférence, après avoir balayé l'ensemble du tableau d'observations et calculé les nombres $m_{k_1,\ldots,k_L}$. Il faut alors les répartir de la manière la plus égale possible entre les cellules concernées.

Références

[MMPR] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN 2-9521458-0-6, ISSN 1767-1175. Mars 2004.

[NMP] Bernard Beauzamy : Nouvelles Méthodes Probabilistes pour l'évaluation des risques. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN 978-2-9521458-4-8. ISSN 1767-1175, avril 2010.

[AMW] Bernard Beauzamy : Archimedes' Modern Works. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN 978-2-9521458-7-9, ISSN 1767-1175. Relié, 220 pages, août 2012.

Table des matières