2019 ■ Vol. 15 ■ no. 3

Suppléer aux données manquantes par maximum de vraisemblance : une application à des variables de Bernoulli
Making up for missing data by maximum likelihood estimation: an application to Bernoulli variables

Louis Laurencelle [a], ✉

[a]Université du Québec à Trois-Rivières

Abstract ■ Statistical parameter estimation from incomplete or lacunar data series is an oft-encountered issue in real settings, an issue for which the user has at his disposal a handful of solutions, from simple and multiple imputation to substitution of an average value, winsorization, and notably least squares estimation (LSQ, or MC in the article) and maximum likelihood estimation (ML, or MV in the article). LSQ and ML allow to fill in a gap in the series by an enlightened and precise estimation of the missing information, a feat that none of the other methods approaches. This advantage of LSQ and ML over the other less appropriate and precise methods is tied up with their drawback: one must know explicitly the probability function of the variable at stake. LSQ and ML estimates are frequently but not always equal and, if LSQ estimation is best suited to and well documented for real variable distributions, it is much less suitable for integer variables. The present article explores ML estimation under conditions of missing data for a few instances of integer Bernoulli variables, namely the Binomial, Geometric, Pascal (or Negative Binomial), Constrained Pascal, and Poisson distributions. Examples with calculations and tables are provided.

Keywords ■ Maximum likelihood estimation, Bernoulli variables, Binomial, Geometric, Pascal, Constrained Pascal, Poisson, Missing data, Truncated distribution .

✉ louis.laurencelle@gmail.com

LL: na

10.20982/tqmp.15.3.p168

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Introduction

Quel chercheur n'a pas, dans sa carrière, été confronté à une situation où son système de mesure lui a fait défaut, où il s'est vu placé devant une série de données incomplète, compromettant son expérimentation? Dans certains cas malheureux, force lui est de renoncer, corriger la situation s'il le peut, puis recommencer. Dans d'autres cas cependant, des solutions existent, solutions tablant sur les données disponibles et permettant de les compléter d'une manière ou d'une autre. Par exemple, si la série tronquée est statique, c.-à-d. si les données perdues sont de même contexte que celles disponibles, on peut les remplacer par la valeur moyenne, en y ajoutant une variabilité congrue si c'est indiqué, ou par l'une ou l'autre des données présentes, choisie au hasard. Dans le cas d'une série interrompue (parce que débordant le temps alloué ou excédant la capacité du système de mesure, ou par une panne du système), la solution de "winsorisation", un pis-aller inspiré de Charles P. Winsor (voir Dixon, 1960), convient parfois, en remplaçant chaque donnée manquante par la donnée extrême de la série, selon sa place ou sa valeur, selon le cas. Dans un contexte multivarié comprenant deux ou plusieurs séries de données, l'imputation

simple ou multiple (Little & Rubin, 1987) peut aussi être envisagée.

Par ailleurs, dans certaines situations, le chercheur a pour but premier d'estimer la valeur d'un paramètre d'intérêt, et il met en place un dispositif, une expérimentation, lui permettant de produire des données qui lui sont mathématiquement associées : ce sera par exemple le niveau typique ou moyen d'une grandeur, la probabilité sous-jacente à une performance mesurée, etc. En voici trois exemples préliminaires.

Exemple 1 – *Estimer le paramètre "μ" d'une population normale à partir d'une série statistique complète.* Le chercheur obtient N mesures dont il suppose qu'elles se distribuent selon la loi normale, c.-à-d. que chaque donnée x obéit à la fonction de densité $\phi(x;\mu,\sigma^2)$ telle que :

$$\phi(x) = \frac{\exp(-(x-\mu)^2/(2\sigma^2))}{\sqrt{2\pi}},$$

et il veut en estimer la valeur μ, caractéristique de la série. La probabilité ponctuelle (ou densité) de chaque observation X_i étant $\phi(X_i)$ et les N observations réputées indépendantes, leur probabilité conjointe est $\prod_{i=1}^{N}(X_i)$, le logarithme de cette expression devenant :

$$\ln \prod_{i=1}^{N}(X_i) = -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) - \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{2\sigma^2}.$$

L'estimation optimale de $\hat{\mu}$ dépend ici, comme on voit, du dernier terme de l'expression logarithmique, le seul terme où μ apparaît. Il s'agit alors de trouver la valeur qui correspond à la probabilité maximale, c.-à-d. dans l'expression logarithmique, la valeur minimale de $\sum_{i=1}^{N}(X_i-\mu)^2/\sigma^2$. La première dérivée de cette fonction de μ étant -2 $\sum_{i=1}^{N}(X_i-\mu)/\sigma^2$, l'extremum de la fonction se situe à -2$\sum_{i=1}^{N}(X_i-\hat{\mu}) = 0$, d'où $\sum X = N \times \hat{\mu}$ et $\hat{\mu} = \sum X/N = \bar{X}$. L'estimateur du maximum de vraisemblance (ou MV) de μ, la première caractéristique de la loi normale, coïncide donc avec la moyenne arithmétique.[1]

Exemple 2 – *Estimer le paramètre "μ" d'une population normale à partir d'une portion sélectionnée de la série statistique.* Par contraste avec la situation quelque peu banale rapportée dans l'exemple précédent, un cas plus complexe peut se présenter, à solution moins courante et requérant un procédé d'estimation moins banal. Un second et dernier exemple utilisant une variable continue X et la loi normale servira d'illustration.

Un chercheur universitaire qui mène une étude d'observation détaillée sur le régime de vie et la santé des étudiantes de premier cycle dans son institution est intéressé à connaître la taille et le poids moyen de cette population, mais son échantillon ne comporte que 10 participantes. Par bonheur, il repêche dans la bibliothèque de la faculté le rapport d'une étude faite l'année précédente sur les facteurs d'obésité des étudiantes, rapport qui produit quelques statistiques pertinentes. Dans ladite étude, on avait d'abord mesuré (poids et taille) 500 personnes, puis sélectionné et recruté les 50 étudiantes de poids plus élevé : les auteurs ne présentent que les statistiques de ce groupe, soit $\bar{X} = 70,14$ et $s_X = 2,71$ kg pour le poids. Ces données, rappelons-le, caractérisent le poids des 50 étudiantes les plus lourdes parmi les 500 mesurées : le poids moyen des 500, un estimateur plus juste du poids typique de la population, est logiquement plus bas. Comment procéder pour inférer ce poids moyen?

Le poids comme la taille se distribuent 'normalement' dans une population homogène, le modèle normal pouvant nous servir pour effectuer l'estimation voulue. Nous avons ici en main les 50 plus hautes valeurs d'une série en comportant 500, soit les 10% supérieures de la population. Soit z distribué selon la densité normale standard $\phi(z)$, la portion des 10% supérieurs est définie par $\int_c^{\infty}\phi(z)\,dz = 0,10$, une table de la loi normale fournissant tout de go la borne $c \approx 1,2816$. Soit les moments m_r de cette sous-variable $c \leq z < \infty$, $m_r = \int_c^{\infty} z^r\phi(z)\,dz$, nous trouvons facilement[2] $m_0 = 1-\Phi(c)$, où $\Phi(c)$ est l'intégrale de $\phi(z)$ pour z de $-\infty$ à c, où $m_1 = \phi(c)$ et $m_2 = 1 + c\cdot\phi(c)/m_0$. Ainsi, utilisant $\phi(c) \approx 0,1755$, la valeur moyenne des 10% supérieures d'une loi normale standard sera $\Delta_c = m_1/m_0 = \phi(c)/(1-\Phi(c)) \approx 1,755$, sa variance, $\sigma^2 = 1+c\cdot\Delta_c-(\Delta_c)^2 \approx 0,1691$ et $\sigma \approx 0,411$. La quantité Δ_c est aussi nommée différentiel de sélection (Laurencelle, 2016) et quantifie l'écart standardisé de la valeur moyenne après sélection d'une moyenne d'une portion sélectionnée des données extrêmes. Une estimation plus précise pour une taille d'échantillon finie, ici $N = 500$, donnerait $\Delta_c(500) \approx 1,750$ et $\sigma \approx 0,408$ (voir aussi Burrows, 1972, 1975).

Donc, afin d'estimer le poids moyen et l'écart-type de la population, on obtient d'abord $\hat{\sigma}_X = s_X/\sigma_z = 2,71/0,411 \approx 6,59$ kg, puis $\hat{\mu}_X = \bar{X} - \hat{\sigma}_X\cdot\Delta_c \approx 70,14 - 6,59\times 1,755 \approx 58,57$ kg: telles sont les valeurs paramétriques du poids estimées pour la population d'étudiantes concernée.

Exemple 3 – *Estimer le paramètre π d'une série d'essais de Bernoulli arrêtée prématurément.* Le troisième et ultime exemple préliminaire porte cette fois sur une vari-

[1] Comme le montre la dernière partie du processus d'estimation, $\bar{X}$ coïncide aussi avec l'estimateur des moindres carrés, étant la valeur $C = \bar{X}$ qui minimise la somme $\sum(X_i - C)^2$. Il est intéressant de rappeler que l'estimateur MV de la variance σ^2 est $\sum(X_i-\mu)^2/N$, comparativement à l'estimateur MC plus connu, $\sum(X_i-\mu)^2/(N-1)$, lequel a la propriété d'être non biaisé *en espérance*.

[2] Noter pour ce faire que $\phi'(x) = -x\cdot\phi(x)$ et $\phi''(x) = (x^2-1)\cdot\phi(x)$, où $\phi'(x)$ et $\phi''(x)$ sont respectivement la première et deuxième dérivée de la fonction $\phi(x)$, d'où on voit, par exemple, que $\phi(x)$ est la primitive de $-x\cdot\phi(x)$.

able de Bernoulli, c.-à-d. ici une variable discrète binaire, la mesure d'un événement pouvant être codée 0 ou 1 : l'événement présente respectivement deux états exhaustifs et mutuellement exclusifs, tels "Succès" et "Échec", "Droite" et "Gauche", "Symptomatique" et "Asymptomatique", etc. Nous entrons ici dans le vif de l'article, soit l'estimation du paramètre à partir d'une série numérique tronquée ou incomplète. Le paramètre-clé, noté π, indique la probabilité de produire le résultat "1" à chaque événement. L'exemple suivant concerne un cas particulier de ce processus, repris en détail plus bas : il s'agit du numéro n d'essai, ou d'événement, auquel on aura accumulé r succès.

Pour illustration, prenons un examen scolaire passé sur ordinateur, grâce auquel l'écran présente à l'élève une suite de petits problèmes d'arithmétique. À chaque problème, l'élève doit désigner la réponse correcte parmi un choix de 3 réponses : l'examen cesse à l'atteinte de 10 réponses correctes ou après le 20^{e} essai. Le délai de réponse accordé à chaque essai est d'une minute, après quoi l'ordinateur attribue au hasard l'une des 3 réponses proposées. Les résultats des 30 élèves, c.-à-d. les numéros d'essai de leur réussite, sont, une fois classés :

11 12^3 13^2 14^2 15^2 16^2 17^4 18^3 19 20^2 $(>20)^8$

1 élève (score 11) a obtenu 10 bonnes réponses dès le 11^{e} essai, en n'en ratant qu'une, et les 8 élèves de queue n'ont pas donné 10 bonnes réponses dans les 20 premiers essais. Quelle est la valeur π qui caractérise la capacité de ce groupe d'élèves, incluant l'information partielle présente dans les "données manquantes", et comment interpréter les résultats de cet examen? La réponse, qui passe par le recours à la loi de Pascal (simple), sera explicitée plus loin.

Dans les sections qui suivent, nous examinerons plus en détail cinq contextes différents, correspondant à cinq lois de probabilités discrètes, les lois binomiale, géométrique, de Pascal, de Pascal restreinte et de Poisson, toutes fondées sur le paramètre π (où $0 < \pi < 1$). Dans chaque cas, les estimateurs MV (maximum de vraisemblance) et MC (moindres carrés) seront fournis et, surtout, nous détaillerons un exemple d'estimation MV pour données incomplètes dans le but d'outiller le lecteur à cette fin. Plusieurs ouvrages documentent de façon compétente et claire l'estimation statistique, par exemple Johnson, Kotz et Balakrishnan (1994, 1995) et le fameux Kendall et Stuart (1979) aux pages 1 à 108. Rohatgi (1976) donne un aperçu succinct mais rigoureux de la question. Et une documentation complète des lois discrètes univariées, incluant nos cinq lois recensées, se trouve Johnson, Kotz, and Kemp (1992).

Une image, qui vaut mille mots, illustrera l'opération d'estimation par maximum de vraisemblance à partir d'un exemple très simple. Supposons que nous avons un dé à 6 faces dont nous voulons estimer empiriquement la probabilité π d'obtenir par chance la face "1". Nous lançons le dé en l'air pour $n = 10$ essais et obtenons $x = 2$ fois la face "1".

Si la probabilité cherchée, $\pi = p(face = 1)$ était près de 1, nous aurions obtenu vraisemblablement plus que 2 fois la face "1", tout comme au contraire si π était 0. En fait, sur n expériences, la probabilité d'obtenir x succès (dans des conditions homogènes et indépendantes) est de $p(x|n, \pi) = p(x) = {}_nC_x \cdot \pi^x(1-\pi)^{n-x}$; nous y revenons à la section suivante.[3]

Nous avons ici $n = 10$ et $x = 2$; il nous faut trouver la valeur de π la plus appropriée, celle en vertu de laquelle la probabilité de notre résultat $x = 2$ sera maximale, c.-à-d. la plus vraisemblable. Le lecteur admettra aisément que, si $\pi = 0$, nous aurions $p(x = 2) = 0$ et notre résultat observé $x = 2$ serait invraisemblable ; c'est aussi le cas si $\pi = 1$, qui aurait plutôt donné lieu à $x = 10$ "faces". La valeur de π cherchée se situe donc entre 0 et 1 : la figure 1 permet de visualiser la fonction ${}_nC_x \cdot \pi^x(1-\pi)^{n-x}$ qui décrit cette relation entre $p(x)$ et π.

Le mode de la fonction, c.-à-d. la vraisemblance[4] maximale de π, égale à peu près 0,302, et ce mode correspond précisément à la valeur $\pi = 0,2$, laquelle est ici l'estimation MV pour le paramètre à l'étude. L'estimation directe par MC est ici la même, plus simplement, x/n ou $2/10 = 0,2$. Ce cas, très simple, admet une solution algébrique MV facile : voir la section suivante. Cette simplicité disparaît toutefois dans les exemples plus réalistes à venir.

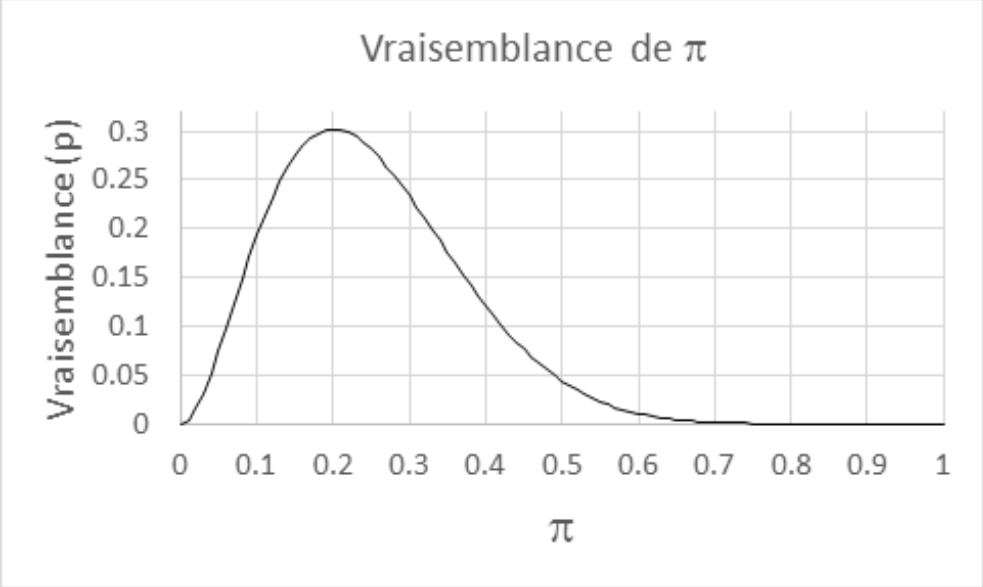
Les deux approches, MV et MC, donnent ici, comme souvent, le même résultat; elles ne le font pas toujours et, dans certains cas comme lorsque des données sont indisponibles, seule l'estimation MV peut nous tirer d'affaire.

Le lecteur, pour le cas présent, aura supposé que la *probabilité* a priori d'obtenir la face "1" d'un dé honnête à 6 faces est de 1/6 ou $\approx 0,1667$. En effet, si le dé est honnête et le test bien mené, la *proportion* x/n va tendre vers la valeur 1/6 à mesure que le nombre d'essais n est augmenté. Chaque accroissement de n donnera alors lieu à une fonction de vraisemblance de plus en plus pointue par comparaison à celle apparaissant à la figure 1, avec un mode glissant vers la valeur d'abscisse 0,166... À l'infini, ladite fonction sera une verticale de mode 1, pointant di-

[3] La notation ${}_nC_x$ indique le nombre de combinaisons de n objets pris x à la fois et correspond à $n!/(x! \cdot (n-x)!)$; d'autres formes de l'opérateur sont $C(n, x)$ ou $\binom{n}{x}$.

[4] On parle ici de "vraisemblance" malgré que le calcul opéré en soit un de probabilité, ceci parce que, au sens strict, la probabilité définit la condition causale en vertu de laquelle le phénomène étudié varie (c.-à-d. le comportement du phénomène dépend de sa fonction de probabilité) tandis qu'on cherche ici la valeur de probabilité la plus forte qui a pu produire le phénomène (le résultat) observé.

Figure 1 ■ Fonction de vraisemblance (binomiale) de l'occurrence de 2 succès en 10 essais.



rectement sur la valeur d'abscisse 1/6.

Notation. La fonction de vraisemblance, qu'on peut noter $u(\pi) = p(f, \pi)$ dans le cas simple ou $u(\pi) = \prod_j^n p_j\,(f_j, \pi)$ dans le cas composé, sert à repérer le mode $\hat{\pi}$ tel que $u(\hat{\pi}) = \max u(\pi)$. Équivalemment et par commodité de calcul, on utilise le logarithme de la fonction tel que $LV(u) = -2 \cdot \ln[u(\pi)]$ [5] qu'il s'agit alors de minimiser, c.-à-d. de rapprocher de 0. Noter que, à l'occasion, la fonction de vraisemblance composée est ramenée à l'unité, c.-à-d. à une valeur de probabilité simple plutôt que sous sa forme d'un produit des probabilités individuelles concernées : il s'agit alors d'une moyenne géométrique, obtenue comme $u(\pi)^{1/N}$, où $u(\pi) = p(f_1, \pi) \cdot p(f_2, \pi) \cdot \ldots \cdot p(f_k, \pi)$ et $N = \sum f$.

Enfin, les calculs rapportés plus bas sont précis à la décimale donnée et selon la précision des calculs qui les précèdent. Par exemple, si $x = \sqrt{2} \approx 1,414$, alors $x^2 = 1,99940$.

Loi binomiale

Variable: La valeur x est un entier positif entre 0 et n et dénote le nombre de succès remportés en n essais à probabilité individuelle π.

Fonction de masse: $p(x : n, \pi) = {}_nC_x \pi^x \cdot \omega^{n-x}$;

Fonction de répartition: pas de forme connue.

Règle de succession: $p(x) = p(x-1) \cdot \pi/\omega \cdot (n+1-x)/x$

Estimation simple: La loi binomiale offre une solution analytique simple pour l'estimation du paramètre π, solution qui passe par une simple dérivée du logarithme de la fonction de masse. Soit $\ln(p) = \ln {}_nC_x + x \cdot \ln(\pi) + (n-x) \cdot \ln(\omega)$. La première dérivée selon π, $d\ln(p)/d\pi = [0 + x \cdot (1/\pi) + (n-x) \cdot (-1/\omega)]/[\pi(1-\pi)]$, est égale à 0 pour la valeur $\pi = x/n$, soit la proportion (ou moyenne) observée.

De là, $\pi_{MV} = x/n$ pour une expérience, et $\sum x_i/(N \cdot n)$ dans le cas composé de N expériences $B(n, x_i)$ homogènes où $\sum x_i / \sum n_i$ à partir de N expériences $B(n_i, x_i)$ inhomogènes. Les mêmes résultats découlent d'un calcul MC.

Moments: $E(x) = n \cdot \pi$, $\sigma^2 = n \cdot \pi \cdot \omega$.

Qu'arrive-t-il maintenant si des données sont manquantes, si la série de mesures n'a pas été complétée?

Loi binomiale tronquée: Les données détaillées de fréquence pour certaines valeurs de x sont manquantes, seule leur fréquence totale est disponible.

Dans l'exemple traité plus bas, comportant n occasions ou essais, chaque protocole binomial est arrêté après le $T - 1^{\text{e}}$ succès, de sorte que pour les valeurs $x = T$ à $x = n$, seul le nombre de protocoles est connu. Nous disposons donc des valeurs x individuelles dans l'intervalle $0..T-1$, permettant les calculs correspondants de $p(x)$. Quant à l'intervalle supérieur défini par $x \geq T$, nous devons considérer globalement $p_c(T) = \sum_{x=T}^{n} p\,(x)$. Noter

[5] La forme "$-2\ln(p)$" est de référence classique, la probabilité p donnée en argument générant ainsi une variable χ^2 à $\nu = 2$ degrés de liberté. La notation LV tient lieu ici du "logarithme de vraisemblance", une valeur positive à propos de laquelle il s'agit de trouver la valeur du paramètre-clé qui minimise LV (maximisant ainsi la probabilité simple ou conjointe correspondante.

que le manque de données, ou la troncature, peut apparaître n'importe où dans la série, voire advenir sur des zones multiples.

Exemple 4

Données: Le tableau 1 montre $N = 18$ participants, chacun sur $n = 10$ essais binomiaux. Les probabilités binomiales et le nombre de participants associés sont indiqués aux lignes $p(x)$ et $f(x)$, avec π (probabilité d'un succès par essai) et $\omega = 1 - \pi$.

Considérant la série complète des $N = 18$ participants (sans troncature, en utilisant aussi les données grisées), la moyenne arithmétique de x est $\sum(f_x \times x)/N = 104/18 = 5,778$, fournissant une estimation de π_{MC}(par essai) $= \sum(f_x \times x/n)/N = 104/(18 \cdot 10) = 0,578$, valeur qui coïncide ici avec l'estimation MV.

Preuve: $\pi = 0,578$ (correspondant à la moyenne géométrique maximale des $p(x)$, égale à 0,12389) donne LV = $-2\sum f_x \times \ln(p(x)) = 75,1817$[6], tandis que 0,577 et 0,579 donnent des valeurs plus basses, respectivement LV = 75,1821 et 75,1826.

Supposons maintenant que les données détaillées de fréquence pour certaines valeurs de x sont manquantes (et représentées ici dans la section grisée du tableau), seul leur nombre étant disponible. Pour notre exemple, le protocole binomial est indisponible pour les T$\leq x \leq n$ succès : le protocole a été stoppé après le T^{e} succès, ou l'appareil de comptage s'est bloqué au nombre T, ou ces données ont été perdues. Nous disposons donc des x et $p(x)$ pour la zone $x = 0..T-1$; quant à la zone complémentaire x de T à n, elle sera représentée par $p_c(T) = \sum_{x=T}^{n} p\,(x)$, une probabilité globale correspondante.

Ici, les données x individuelles dans la zone supérieure $8 \leq x \leq 10$ sont manquantes, hormis qu'on y trouve 2 protocoles enregistrant 8 succès ou plus. Il s'agit alors de maximiser le produit des probabilités mesurables, soit les $N = 18$ probabilités associées aux valeurs de la variable : d'une part, nous avons 16 probabilités individuelles $p(x)$ dans la zone détaillée, pour x de 0 à 7, et d'autre part la probabilité complémentaire $p_c(T)$, pour les 2 données masquées (la probabilité qu'elles relèvent de la zone concernée est caractérisée par la somme des probabilités associées à cette zone). On doit donc repérer la valeur de π qui maximise $u(\pi)$ ou minimise le LV correspondant, soit

$$u(\pi) = \prod_{x=0}^{\text{T}-1} p(x,\pi)^{f_x} \cdot \left[\sum_{\text{T}}^{n} p(x,\pi)\right]^{N-\sum_{x=0}^{\text{T}-1} f_x}.$$

Noter l'exposant f_x associé à chaque terme de probabilité, notamment le terme composé à droite, la probabilité composée (à partir des probabilités individuelles $\sum p(x)$, $x = T$ à n) étant élevée ici à la puissance 2, correspondant aux 2 données masquées.

Par recherche convergente entre les bornes 0 et 1, on trouve $\pi = 0,564$ (LV = 65,119392), associée à la moyenne géométrique des $p(x)$ de 0,16384 qui à sont tour correspond à la vraisemblance maximale des données de l'exemple[7]. Le lecteur notera que cette valeur estimée ($\pi = 0,564$) est inférieure à celle obtenue pour la série complète ($\pi = 0,578$), la différence émanant de l'influence des 2 valeurs supérieures cachées. Dans le cas de la série complète, les données x_i extrêmes ajoutent 9+10 = 19 au total, alors que l'"ajout" correspondant dans le calcul MV est $2 \times 8,26 = 16,52$[8]: tout se passe comme si ces valeurs $x = 9$ et 10 déviaient vers le haut par rapport à l'information fournie dans les bas résultats rapportés, information telle que produite par l'estimation MV.

Note. Il se trouve différents moyens à la disposition de l'usager pour repérer et fixer la valeur cherchée, notamment en exploitant la fonction `Solveur` du logiciel Excel de Microsoft, voire par essais et erreurs. En général et pour éviter la perte de précision numérique, il est recommandé de procéder par la forme logarithmique du calcul, c.-à-d. par la fonction LV, laquelle permet de préserver un plus grand nombre de chiffres significatifs d'une opération à l'autre.

Loi géométrique

Variable: La valeur n est un entier positif et dénote le numéro d'essai auquel le processus de Bernoulli produit le premier succès, chaque essai ayant probabilité π de produire un succès.

La loi exponentielle $E(b : x)$ fournit une bonne approximation de la loi géométrique, soit $p(x) = exp[-(x - \frac{1}{2})/b]/b$, et $P(x) = 1 - exp[-(x - \frac{1}{2})/b]$, où b est fonction de π. Nous avons trouvé $b \sim 0,5282 \cdot \pi^{-1,278}$, avec $R^2 = 0,983$, la relation entre le b calculé et sa valeur expérimentale obtenant $R^2 = 0,991$.

Fonction de masse: $p(n : \pi) = \pi\omega^{n-1}$;

Fonction de répartition: $P(n) = 1 - \omega^n$

Règle de succession: $p(n) = p(n-1) \cdot \omega$

Estimation: $\pi_{MV} = \pi_{MC} = 1/n$, ou $\pi = N/\sum n_i = 1/\bar{n}$ dans le cas de plusieurs (N) expériences.

Moments: $E(n) = 1/\pi; \sigma^2 = \omega/\pi^2$

[6] La valeur donnée $\pi = 0,578$ a été bloquée à 3 décimales (la valeur juste étant $0,5\overline{7} \approx 0,57777...$, la quantité corrélative LV rapportée correspondant à cette valeur simplifiée.

[7] La vraisemblance maximale est donc égale au produit des 18 probabilités en jeu, soit $0,16384^{18}$.

[8] Le calcul correspond en effet à $(\sum_{x=0}^{7}(f_x \cdot x) + Q)/(N \cdot n) = (85 + Q)/180 = \pi$. Pour la série complète, la valeur π obtenue donne $Q = \pi \cdot 180 - 85$, ce qui, pour la série complète ($\pi \approx 0,568$) indique $Q = 19$ et, pour la série tronquée ($\pi \approx 0,564$), $Q = 16,52$.

Table 1 ■ Nombre de succès (x) sur 10 essais binomiaux pour un groupe de $N = \sum f = 18$ participants, avec la probabilité associée à chaque nombre

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	ω^{10}	$10\pi\omega^9$	$45\pi^2\omega^8$	$120\pi^3\omega^7$	$210\pi^4\omega^6$	$252\pi^5\omega^5$	$210\pi^6\omega^4$	$120\pi^7\omega^3$	$45\pi^8\omega^2$	$10\pi^9\omega$	π^{10}
$f(x)$	0	0	1	0	3	5	3	4	0	1	1

Exemple 5

La série de $N = 8$ valeurs n_t suivantes: $n_i = 3, 5, 6, 6, 9, [11, 15, 21]$, a pour moyenne ($\bar{X}$) = 76/8 = 9,5 et l'estimateur MV donne $\pi = 8/76 \approx 1/\bar{X} \approx 0,105$. La chaîne de probabilités correspondantes est $\pi\omega^2, \pi\omega^4, (\pi\omega^5)^2 \cdot \ldots \cdot \pi\omega^{20}$ et leur produit, $\pi\omega^2 \cdot \pi\omega^4 \cdot (\pi\omega^5)^2 \cdot \ldots \cdot \pi\omega^{20}$. La valeur $\pi = N/\sum n_i = 0,105$ donne ici LV = 51,147, contre 51,149 et 51,148 pour $\pi = 0,104$ et 0,106 respectivement.

Si maintenant nous bloquons les données après l'essai $n = 10$ – une procédure courante en recherche et dans la pratique de testing, histoire de ne pas éterniser l'expérience –, nous trouvons 5 expériences bien numérotées, plus 3 arrêtées prématurément, tel qu'indiqué par les crochets dans la série ci-dessus. La probabilité qu'un (premier) succès advienne au T^{e} essai ou plus loin est la somme :

$$Q(T) = p(n \geq T) = \sum_{n=T}^{\infty} \pi \cdot \omega^{n-1} = \omega^T$$

et est appelée fonction de survie. Les valeurs des trois protocoles escamotés sont alors remplacées par cette probabilité globale, appliquée 3 fois.

Prenant $T = 11$, le produit des probabilités pour notre exemple tronqué devient alors $\pi\omega^2 \cdot \pi\omega^4 \cdot (\pi\omega^5)^2 \cdot \pi\omega^8 \cdot (\omega^{11})^3$, le LV trouvé donnant la valeur minimale de 34,762 à $\pi \approx 0,081$.

Par comparaison, l'estimateur basé sur les seules 5 données 'disponibles' serait $\hat{\pi} \approx 0,172$ (correspondant à $\bar{X} = 5,8$), alors que la série winsorisée, obtenue en allongeant la série de trois valeurs 11 fictives, donnerait $\hat{\pi} \approx 0,129$ ($\bar{X}$ = 7,75), notre estimateur μ_{MV} de 0,081 (correspondant à une moyenne déduite $\bar{X}$ de 12,3) étant celui qui s'approche le plus de l'estimateur basé sur la série complète, soit 0,105.

Loi de Pascal

Variable: La valeur n est l'entier positif dénotant le numéro d'essai auquel le processus de Bernoulli produit le $r^{\text{e}} (r \geq 1)$ succès, chaque succès ayant probabilité π.

La loi géométrique est un cas particulier de la loi de Pascal, selon $r = 1$. Cette loi est comparativement peu documentée dans la littérature spécialisée. Elle y est associée à la loi Binomiale négative, dans laquelle la variable n est un nombre réel positif, pas nécessairement entier, et dont elle est un cas particulier.

La loi Gamma $G(b, c : x)$, connue aussi sous le nom de loi d'Erlang, peut servir à approcher la loi de Pascal, en utilisant $p(x) = ((x - \frac{1}{2})/b)^{c-1} \times exp(-(x - \frac{1}{2})/b)/(b \cdot \Gamma(c))$.
Fonction de masse: $p(n : r, \pi) = {}_{n-1}C_{r-1} \cdot \pi^r \cdot \omega^{n-r}$;
Fonction de répartition: pas de forme générale connue[9]
Règle de succession: $p(n) = p(n-1) \cdot (n-1)/(n-r) \cdot \omega$
Estimation: $\pi_{MV} = r/n, \pi_{MC} = (r-1)/(n-1)$
Moments: $E(n) = r/\pi; \sigma^2 = r \cdot \omega/\pi^2$
Note sur les fonctions de répartition $P(n)$ et de survie $Q(n)$. La littérature ne rapporte pas d'expression générale pour la somme d'une séquence de probabilités de cette loi. Une expression à calcul incomplètement déterminé pour cette somme est :

$$1 - P(n-1) = Q(n) = \sum_{x=n}^{\infty} p(n) = \frac{\omega^{n-r}}{(r-1)!} \sum_{x=0}^{r-1} \pi^x \left\{ \sum_{u=0}^{x} c_{r,x,u} n^u \right\}.$$

Selon $r = 3$, par exemple, la formule $Q_{r=3}(n)$ est $\frac{\omega^{n-3}}{(3-1)!}\left\{2 + \pi\left(-6 + 2n\right) + \pi^2\left(6 - 5n + n^2\right)\right\}$. Le tableau 2 présente un sous-ensemble des coefficients $c_{r,x,u}$de la fonction de survie, pour $r = 2$ à 7: la règle générale de calcul des coefficients reste à trouver.[10]

Exemple 6

Référons-nous à un protocole dans lequel chaque participant doit, par une succession d'essais, accumuler $r = 3$ succès; cependant le protocole est stoppé après le 10^{e} essai. Voici les données de $N = 15$ participants fictifs (numéros d'essai final replacés en ordre croissant); celles excédant le 10^{e} essai, réputées indisponibles, sont inscrites entre crochets :

$n = 5 \quad 5 \quad 7 \quad 8 \quad 8 \quad 8 \quad 9 \quad 10 \quad 10 \quad [12\ 15\ 19\ 19\ 25\ 32]$

ou, en notation raccourcie:

$$n = 5^2 \quad 7 \quad 8^3 \quad 9 \quad 10^2 \quad [12\ 15\ 192\ 25\ 32].$$

[9] Référence Wiki : https://en.wikipedia.org/wiki/Negative_binomial_distribution#Cumulative_distribution_function

[10] Un calcul de la fonction de répartition $P(n)$, ou $Q(n)$, est possible via une fonction Bêta incomplète, qu'il faut ensuite inverser pour estimer π, solution pour laquelle il existe aussi des approximations; voir Johnson et al., 1992, p. 209-213.

Table 2 ■ Coefficients de la fonction de survie $Q(n)$ de la loi de Pascal pour $r = 2$ à 7 succès

		u						
r	x	0	1	2	3	4	5	6
2	0	1						
	1	-2	1					
3	0	2						
	1	-6	2					
	2	6	-5	1				
4	0	6						
	1	-24	6					
	2	36	-21 3					
	3	-24	26	-9	1			
5	0	24						
	1	-120	24					
	2	240	-108	12				
	3	-240	188	-48	4			
	4	120	-154	71	-14 1			
6	0	120						
	1	-720	120					
	2	1800	-660	60				
	3	-2400	1480	-300	20			
	4	1800	-1710	595	-90	5		
	5	-720	1044	-580	155	-20	1	
7	0	720						
	1	-5040	720					
	2	15120	-4680	360				
	3	-25200	12840	-2160	120			
	4	25200	-19140	5370	-660	30		
	5	-15120	16524	-7050	1470	-150	6	
	6	5040	-8028	5104	-1665	295	-27 1	

Note. $Q(n) = 1 - P(n-1) = \sum_{x=n}^{\infty} p\,(n) = \frac{\omega^{n-r}}{(r-1)!} \sum_{x=0}^{r-1} \pi^x \left\{\sum_{u=0}^{x} c_{r,x,u} n^u\right\}$, où $\omega = 1 - \pi$.

Noter que le nombre moyen d'essais requis par les 15 participants est de 12,80, fournissant pour π l'estimation $(3-1)/(\bar{X}-1) \approx 0,156$ par MC et $3/\bar{X} \approx 0,234$ par MV. Dans le cas tronqué maintenant, le protocole ayant été stoppé après le 10e essai, la seule information disponible pour les participants concernés est que 6 ont (virtuellement) réussi à un essai quelconque au-delà du 10e. La solution à trouver consiste donc à déterminer la valeur de π qui maximise le produit (ou minimise le LV) de l'expression suivante :

$$p(5)^2 \cdot p(7) \cdot p(8)^3 \cdot p(9) \cdot p(10)^2 \cdot [1 - \sum_{n=3}^{10} p\,(n)]^6.$$

La probabilité globale inscrite entre crochets, équivalente à la fonction de survie $Q(11)$, peut être calculée explicitement. Le lecteur peut aussi recourir aux séries de coefficients données au tableau 2, pour $r = 2$ à 7 succès. Pour $r = 3$, l'expression vue plus haut s'applique, soit :

$$Q_3(n) = \frac{\omega^{n-3}}{(3-1)!} \left\{2 + \pi\,(-6+2n) + \pi^2\left(6 - 5n + n^2\right)\right\},$$

en utilisant ici $n = 11$.

À chaque itération du calcul, que ce soit par essais et erreurs ou grâce à un algorithme de convergence, les 5 probabilités individuelles (ou leurs logarithmes) de même que la somme indiquée entre crochets ou la fonction Q(11) correspondante doivent être recalculées. Après minimisation du critère LV à la valeur 55,04644, l'estimation π_{MV} correspondante est 0,276.

Solution du problème de l'Exemple 3. Une fois défini le terme de probabilité $p(n : r = 10, \pi)$ associé à chaque résultat n observé, p. ex. $p(12) = {}_{12-1}C_{10-1} \cdot \pi^{10} \cdot \omega^{12-10}$, et l'expression établie pour les valeurs débordantes, à savoir $p(n > 20) = 1 - \sum_{n=10}^{20} p\,(n)$, la probabilité globale attachée aux résultats des 30 participants peut être cal-

culée, la forme logarithmique (LV) étant privilégiée. Le lecteur vérifiera que la valeur minimale LV $\approx 139,12$ est atteinte pour $\hat{\pi} \approx 0,567$, soit la probabilité que l'élève typique trouve la bonne réponse parmi les 3 proposées pour chaque problème. Noter que cette valeur $\hat{\pi}$ n'excède pas généreusement la probabilité $\frac{1}{3}$ de trouver la bonne réponse par hasard. Noter aussi que, pour les 8 élèves n'étant pas arrivés à produire leurs 10 réponses correctes avant le 21e problème, l'estimateur MV de leur "capacité" individuelle est inférieur à $r/n = 10/20 = \frac{1}{2}$, et il pourrait même flirter avec $\frac{1}{3}$, correspondant à la réponse au hasard.

Loi de Pascal restreinte

Variable: La valeur n est un entier positif qui dénote le numéro d'essai auquel le processus de Bernoulli produit pour le première fois r succès parmi les k plus récents essais, où $r \leq k$, chaque succès ayant probabilité π. La fonction de masse étant notée $p(n : k, r, \pi)$, la loi géométrique $(k = r = 1)$ et la loi de Pascal $(k \to \infty)$ s'en trouvent être des cas particuliers.

Cette loi a été identifiée et partiellement définie pour la première fois dans Laurencelle (1983). La mathématique pour l'ensemble des cas de cette loi n'est pas encore complétée, de sorte que ni les fonctions de masse et de répartition ni la règle de succession ne sont généralement établies.

Une formule générale 'aveugle' pour la fonction de masse de cette loi serait :

$$\begin{aligned} p(n : k, r, \pi) &= p(n) \\ &= \sum_{u \geq r} c_u \cdot \pi^u (1-\pi)^{n-u} (n \geq r), \end{aligned} \qquad \text{(A)}$$

les coefficients c_u restant à déterminer selon le cas.

La loi comporte deux cas spéciaux, soit la loi des succès consécutifs, $p(n : r, r, \pi)$ et la loi restreinte pour $r = 2$ succès, $p(n : k, 2, \pi)$, ces deux cas ayant des solutions spécifiques et complètes. Le cas général est présenté plus bas.

Cas spécial - loi des succès consécutifs $p(n : r, r, \pi)$: la variable n dénote le numéro d'essai auquel le processus produit pour la première fois r succès consécutifs. Cette loi, plus récemment rééditée sous le nom de loi géométrique d'ordre r,[11] a été initialement définie par Abraham de Moivre en 1758 (The doctrine of chances, problème LXXXVII) : "To find the probability of throwing a chance assigned a given number of times without intermission, in any given number of trials".

Outre la solution de de Moivre, la probabilité $p(n) = p(n : r, r, \pi)$ admet d'autres solutions, dont celles de Mood (1940) et de Bradley (1968), dont nous nous inspirons (Laurencelle, 2012). La fonction de répartition, sous forme récursive, est donnée par Uspensky (1937).

Fonction de masse:

$$\begin{aligned} p(n) &= 0 \text{ si } n < r \\ &= \pi^r \text{ si } n = r \\ &= \omega \cdot \pi^r \text{ si } r < n < 2r \\ &= \left(1 - \sum_{i=2k}^{n-1} (\pi^{i-r} \omega^{n-1-i}) \times \right. \\ &\quad \left. \left(\sum_{j=1}^{(i-r)/r} (-1)^{j-1} \binom{n-i}{j} \binom{n-r-1-r\cdot j}{n-1-i} \right) \right) \omega \pi^r \\ &\quad \text{si } n \geq 2r. \end{aligned}$$

Nous référant à la formule générale (A) ci-dessus appliquée à ce cas, il appert que les coefficients c_u présentent une structure simple et utile mais incomplètement définie, soit:

$$c_u = (1+1)^t_{(u)}, t = n - r - 1, 0 \leq u \leq t, r+1 \leq n \leq 2r,$$

la variable n naviguant ente $r+1$ et $2r$. Les coefficients c_u pour n de $2r+1$ et plus, utilisés dans l'expression (A), sont à encore à déterminer.

Par exemple, pour $r = 5$ et $n = 7$, nous avons $t = 7 - 5 - 1 = 1$ et développant $(1+1)^1_{(u)} = 1,1$ et $p(7) = 1 \cdot \pi^5 \omega^2 + 1 \cdot \pi^6 \omega$. Pour $n = 10$, $t = 4$, d'où $c_u = (1+1)^4_{(u)} = 1,4,6,4,1$ et $p(10) = 1 \cdot \pi^5 \omega^5 + 4 \cdot \pi^6 \omega^4 + 6 \cdot \pi^7 \omega^3 + 4 \cdot \pi^8 \omega^2 + 1 \cdot \pi^9 \omega^1$. Les paliers u supérieurs perdent des unités pour $n > 2r$: leur détermination reste à trouver.

Fonction de répartition: la fonction de répartition s'obtient par récurrence, soit : $P(n) = 0$ pour $0 \leq n < r$ $P(r) = \pi^r$, $P(n) = P(n) + [1 - P(n-k-1]\,\omega\pi^r$ pour $n > r$

Règle de succession: $p(n) = [1 - P(n-k-1]\,\omega\pi^r$ pour $n > r$

Moments:

$$E(n) = \frac{1-\pi^r}{\omega\pi^r};$$

$$\sigma^2 = \frac{1 - \pi^{(2r+1)} - (2r+1)\cdot\omega\pi^r}{(\omega\cdot\pi^r)^2}.$$

Estimation: Outre l'estimation MV (voir plus bas), on peut utiliser l'inversion numérique des moments, notamment via l'explicitation de l'approximation $E(n) \approx \bar{X}$.

[11] La loi géométrique d'ordre r décrit en fait la distribution du numéro d'essai (ou du nombre d'essais) requis pour obtenir la x^e occurrence d'une suite de r succès. Le premier effort moderne pour étudier cette loi remonte vraisemblablement à Uspensky (1937) et Mood (1940), réactualisé par Feller (1968), et c'est Georghiu et Philippou (1983) qui ont relancé la recherche et baptisé la loi sous son nom actuel.

Cas spécial - loi de Pascal restreinte pour 2 succès, $p(n : k, 2, \pi)$: la variable n dénote le numéro d'essai auquel deux succès ($r = 2$) sont rapportés pour la première fois dans les k plus récents essais.

Fonction de masse: La fonction de masse pour cette loi a été trouvée (sous différentes formes) par Koutras (1996) et Laurencelle (1998). Utilisant la formule générale (A) indiquée en début de section, les coefficients c_u obéissent au modèle suivant :

$$c_2 = max[0, min(n-1, k-1)]$$
$$c_3 = \binom{n-(k-1)-1}{2} - \binom{n-2(k-1)-1}{2}$$
$$c_4 = \binom{n-2(k-1)-1}{3} - \binom{n-3(k-1)-1}{3}$$
$$c_5 = \binom{n-3(k-1)-1}{4} - \binom{n-4(k-1)-1}{4}$$

et ainsi de suite. Noter que $\binom{n}{x} = {}_nC_x = 0$ si $n < x$. Par exemple, pour $k = 4$ et $n = 6$, c_2 donne 3 et $c_3 = 1$, d'où $p(6) = 3\pi^2\omega^4 + 1\pi^3\omega^3$, alors que pour $n = 12$, nous obtenons $c_2 = 3$, $c_3 = 18$ et $c_4 = 10$, d'où $p(12) = 3\pi^2\omega^9 + 18\pi^3\omega^8 + 10\pi^4\omega^7$.

Koutras (1996) propose un calcul par récursion, soit: $p(n) = p(n-1)\cdot\omega + p(n-k)\cdot\pi\cdot\omega^{k-1}$, à partir des valeurs $p(n) = (n-1) = \omega^{n-2}\pi^2$ pour $2 \le n \le k$.

Fonction de répartition: pas de forme connue

Moments:

$$E(n) = \frac{2-\omega^{k-1}}{\pi(1-\omega^{k-1})};$$

$$\sigma^2 = \frac{2\omega + \omega^{k-1}\left((2k-1) - (2k+1)\cdot\omega + \omega^k\right)}{\left(\pi(1-\omega^{k-1})\right)^2}.$$

Estimation: Outre l'estimation MV (voir plus bas), on peut utiliser l'inversion numérique des moments, notamment via l'explicitation de l'approximation $E(n) \approx \bar{X}$.

Cas général: Il n'existe pas d'expression algébrique complète de la fonction de masse pour le cas général de la loi de Pascal restreinte, laquelle fonction obéit toutefois à la forme (A), soit: $p(n : k, r, \pi) = \sum_{u \ge r} c_u \pi^u (1-\pi)^{n-u}$.

Les coefficients sont à trouver par énumération, par exemple:

- coefficients c_u pour $k = 4$, $r = 2$ (tel qu'illustré aussi à la sous-section précédente) : $(n : c_u) = (2 : 1), (3 : 2), (4 : 3), (5 : 3), (6 : 3-1), (7 : 3-3), (8 : 3-6), (9 : 3-9), (10 : 3-12-1)$, etc.
- coefficients c_u pour $k = 6$ et $r = 3$ (cette fois, sans recours connu autre que l'énumération) : $(n : c_u) = (3 : 1), (4 : 3), (5 : 6), (6 : 10), (7 : 10), (8 : 10-4), (9 : 10-11-1), (10 : 10-20-6), (11 : 10-30-20)$.

Par exemple, pour $k = 6$, $r = 3$ et $n = 10$, le code (10:10-20-6) se développe comme $10\pi^3\omega^7 + 20\pi^4\omega^6 + 6\pi^5\omega^5$. Aussi, pour $k = 4$ et $r = 2$, nous avons $p(2) = 1\pi^2\omega^0$; $p(3) = 2\pi^2\omega^1$;... $p(7) = 3\pi^2\omega^5 + 3\pi^3\omega^4$;... $p(10) = 3\pi^2\omega^8 + 12\pi^3\omega^7 + 1\pi^4\omega^6$.

Fonction de répartition: aucune expression générale connue pour cette loi.

Moments: Il n'y a pas d'expression générale connue pour les moments de cette loi (sauf les cas particuliers et les deux cas spéciaux présentés ci-dessus). Laurencelle (1983, 2012) présente une méthode de calcul basée sur le recours à une chaîne de Markov.

Estimation: On ne trouve pas d'estimateur MC pour le paramètre π de cette loi. Quant à l'estimation par maximum de vraisemblance, on peut l'obtenir par inversion de la chaîne de probabilités associée à une réalisation ou quelques réalisations de la loi : les exemples offerts dans la sous-section suivante illustrent la procédure.

Table de coefficients et son usage: Dans le matériel supplémentaire joint à l'article, le lecteur trouvera une série de tableaux 3, énumérant les coefficients c_u utiles à compléter le calcul de probabilité grâce à l'expression (A) ci-dessus. Ces coefficients, obtenus par énumération arborescente sur ordinateur, couvrent les domaines paramétriques $r = 2$ à 10 et $k = r$ à $r + 5$, ce pour n allant de $k + 1$ à $k + 10$. Quant aux coefficients appropriés pour $n < k+1$, on les obtient facilement selon : $c_u = 0$ pour $n < r$, tout u; $c_r = {}_{n-1}C_{r-1}$ et $c_{u>r} = 0$ pour $r \le n \le k$.

Illustration 1: Pour le contexte $r = 6$, $k = 9$, $n = 8$, comme $n \le k$, nous avons $c_6 = {}_{8-1}C_{6-1} = 21$ et $c_u = 0$ pour $u > 6$, d'où $p(n = 8) = 21\pi^6\omega^{8-6}$.

Illustration 2: Pour le contexte $r = 6$, $k = 9$, $n = 13$, la table donne 56, 140, 90 et 10 pour $u = 6$, 7, 8 et 9, d'où, simplement, $p(n = 13) = 56\pi^6\omega^{13-6} + 140\pi^7\omega^{13-7} + 90\pi^8\omega^{13-8} + 10\pi^9\omega^{13-9}$.

Exemple 7

Note: L'estimation du paramètre π pour la présente loi ne comporte, à notre connaissance, que deux avenues de solution : par maximum de vraisemblance, ce que notre exemple ci-dessous illustrera, et par l'inversion du calcul des moments (espérance, variance), ce dans les rares cas où ceux-ci sont établis. L'exemple présenté, déjà un peu complexe, ne comporte pas de données manquantes : le cas échéant, le lecteur pourra s'inspirer de la procédure illustrée pour les autres lois, la logique d'estimation par inclusion des probabilités concernées étant la même.

Dans une industrie mécanisée produisant à la chaîne des boîtes contenant chacune 5 composants identiques, la phase d'empaquetage consiste à placer et sceller sous vide les composants dans la boîte dans le délai temporel

prescrit, opération qui peut échouer pour chaque composant, advenant quoi la chaîne doit être temporairement stoppée. La durée d'insertion d'un composant étant d, le temps alloué à la phase d'empaquetage est de $7d$, correspondant au temps virtuellement requis pour traiter 5 + 2 = 7 composants : l'ajout de durée pour 2 composants supplétifs potentiels assure une tolérance temporelle pour le fonctionnement continu de la chaîne. La question est de déterminer quelle est la probabilité d'arrêt de la chaîne (indiquée par une tendance marquée à excéder le temps de complétion alloué (durée de 7 composants) en raison d'un nombre d'échecs trop élevé.

Quatre essais indépendants de l'opération de la phase seule d'empaquetage ont produit les résultats suivants :

$$n = 7, 8, 11 \text{ et } 15,$$

soit les nombres d'essais requis pour caser 5 composants dans une durée $7d$.

Nous avons ici un modèle de loi de Pascal restreinte, de paramètre $r = 5$ (réussites) et $k = 7$ (en 7 essais alloués). Utilisant la formule générale (A) fournie plus haut, les coefficients utiles à l'estimation sont, respectivement (voir tableaux 3, en matériel supplémetaire sur le site web de la revue):

$$\begin{aligned} n &= 7 : 15 \\ n &= 8 : 15 \\ n &= 11 : 15\ 39\ 27\ 3 \\ n &= 15 : 15\ 99\ 273\ 399\ 315. \end{aligned}$$

Soit π, la probabilité de succès par composant; on peut estimer la valeur de π telle qu'elle correspond au maximum de la probabilité associée $p(n : \pi)$.

$$\begin{aligned} \text{Pour } n = 7, p(7) &= 15\pi^5\omega^2, \text{ où } \omega = 1 - \pi \\ n = 8, p(8) &= 15\pi^5\omega^3 \\ n = 11, p(11) &= 15\pi^5\omega^6 + 39\pi^6\omega^5 + 27\pi^7\omega^4 + 3\pi^8\omega^3 \\ n = 15, p(15) &= 15\pi^5\omega^{10} + 99\pi^6\omega^9 + 273\pi^7\omega^8 + \\ &\quad 399\pi^8\omega^7 + 315\pi^9\omega^6. \end{aligned}$$

Chaque essai, avec son résultat "n", peut ainsi donner lieu à une estimation indépendante de π, l'estimation globale, basée sur la combinaison des quatre "probabilités" ci-dessus, représentant la totalité de l'expérimentation.

Les cas pour lesquels $n \leq r+1$ se situent dans une zone compatible avec la loi de Pascal, où l'estimateur du maximum de vraisemblance est r/n. Ainsi, pour les essais $n = 7$ et $n = 8$, les valeurs estimées sont $\hat{\pi}_1 = 5/7 \approx 0,7143$ et $\hat{\pi}_2 = 0,6250$. L'estimation par recherche itérative de π au maximum de vraisemblance fournit $\hat{\pi}_3 = 0,5832$ et $\hat{\pi}_4 = 0,5284$ pour les deux essais suivants. Quant à l'estimation globale, il faut définir une probabilité globale, selon $p_{\text{globale}} = p(7) \cdot p(8) \cdot p(11) \cdot p(15)$, laquelle, maximisée à $\approx 0,06644$ (correspondant à $LV_{min} \approx 21,698$) par le même procédé itératif, résulte en $\hat{\pi}_{globale} \approx 0,6008$.

Stipulant cette valeur $\hat{\pi} = 0,6008$ comme meilleure approximation de π, la probabilité que la chaîne soit retardée devient calculable. La probabilité de non-retard suppose une réussite en $n = 5$ à 7 essais, soit $\sum p(n, \hat{\pi}) = 1\pi^5\hat{\pi}^0 + 5\hat{\pi}^5\omega^1 + 15\hat{\pi}^5\omega^2 = 0,4216$ (où $\omega = 1 - \hat{\pi}$). La probabilité de retard estimée est donc de $1 - 0.4216 = 0,5784$, soit un taux d'arrêt de $1/\pi = 1/0.4216 \approx 2,37$ ($\sigma \approx 1,80$: voir Loi géométrique) ou un peu plus d'une fois sur deux. En conclusion pratique pour cet exemple fictif, mieux vaudrait modifier l'appareillage ou, sinon, allonger la durée de tolérance allouée au-delà $2d$. Une marge ajoutée de $6d$, soit une durée totale de $11d$, fournirait une probabilité $\pi \approx 0,902$, avec taux d'arrêt de 1,11 ($\sigma \approx 0,35$), une valeur beaucoup plus confortable.

Loi de Poisson

Variable: La valeur x, un entier positif, dénote le nombre d'événements produits dans un intervalle (de temps, d'espace, etc.) δ, ou dans un ensemble d'une taille δ donnée, la probabilité π de l'événement étant petite. Posant π_i la probabilité instantanée ou ponctuelle de l'événement visé et δ, la longueur de l'intervalle ou la taille de l'ensemble observé, le paramètre de la loi de Poisson est $\lambda = \mu = \delta = \pi_i$. Cette loi appartient donc à la famille Bernoulli grâce à son paramètre occulte π. La loi de Poisson $p_P(x|\lambda)$ sert d'approximation à la Binomiale $p_B(x|\pi)$ pour π diminuant et n croissant tels que $\pi = \lambda/\delta$ est constant : p. ex. pour $\pi = 0,05$ et $\delta = 100$, $\lambda = 5$, et $p_B(3) \approx 0,13958$ et $p_P(3) \approx 0,14037$.[12]

Noter que l'intervalle (temporel, spatial) entre les événements successifs de type Poisson obéit à la loi exponentielle.

Différentes techniques permettent d'approximer les fonctions de probabilité de cette loi, notamment par la loi continue du khi-deux (Johnson et al., 1992, p. 162 seq.).
Fonction de masse: $p(x : \lambda) = e^{-\lambda} \cdot \lambda^x/x!$;
Fonction de répartition: pas de forme connue
Règle de succession: $p(x + 1) = p(x) = \lambda/(x + 1)$.
Estimation: μ_{MV} et $\mu_{MC} = E(x) \approx \bar{X}$. Noter qu'une inégalité patente entre la moyenne et la variance (voir Moments, plus bas) des valeurs x_i observées récuse la légitimité du modèle Poisson pour ce cas.
Moments: $E(x) = \lambda$; $\sigma^2 = \lambda$

[12] La binomiale ${}_nC_x \cdot \pi^x(1-\pi)^{n-x}$, une fois renotée, devient ${}_nC_x \cdot (\lambda/n)^x \cdot (1-\lambda/n)^{n-x}$. Or, comme $n \gg x$ et $x/n \approx 0$, nous pouvons écrire ${}_nC_x \times (\lambda/n)^x \approx \lambda^x/x!$ et $(1-\lambda/n)^{n-x} \approx (1-\lambda/n)^n$, laquelle expression tend vers $e^{-\lambda}$ pour n croissant, d'où nous obtenons en approximation la loi de Poisson, $e^{-\lambda} \cdot \lambda^x/x!$.

Exemple 8

Soit les statistiques de pannes mensuelles d'une trieuse de courrier, collectées sur 8 mois :

$$x_i = 6, x_2, 3, 9, 5, x_6, 1, 8;$$

les données identifiées par les lettres x_2 et x_6 sont indisponibles mais plus élevées que 9, le système arrêtant de compter après 9 pannes. On cherche ici à estimer le nombre moyen approximatif approprié.

La probabilité d'observer 10 pannes ou davantage étant le complément d'en observer moins de dix, nous avons $p(x \geq 10) = 1 - \sum_{x=0}^{9} p(x)$. La vraisemblance à maximiser sera alors :

$$p(\text{ensemble } x_i) = p(1) \cdot p(3) \cdot p(5) \cdot p(6) \cdot p(8) \cdot p(9) \cdot [p(x \geq 10)]^2.$$

Il s'agit alors de développer la somme indiquée par l'inéquation $p(x \geq 10)$ et l'incorporer dans le produit ci-dessus (ou dans la somme LV correspondante). Cette probabilité est maximisée à $2,468 \times 10^{-9}$ (ou $LV = 39,640$), la valeur ciblée du paramètre étant $\hat{\lambda} = \hat{\mu} \approx 6,78$. La moyenne tronquée, c.-à-d. n'exploitant que les seuls 6 mois à valeurs rapportées, serait de 5,33; avec winsorisation, soit en ajoutant ici deux valeurs "10",[13] l'estimation deviendrait $\hat{\mu} = 6,50$.

Épilogue

Les situations dans lesquelles le chercheur a perdu l'accès à certaines données sont fréquentes, et cela pour toutes sortes de raisons. La suppléance des données manquantes, quand elle est possible, permet souvent de sauver un stock de données et finaliser ainsi une expérimentation par ailleurs valable, ou une évaluation. Quelles sont les conditions propices à un tel sauvetage, pour une série statistique particulière? La première condition est que chaque donnée manquante soit qualifiée, c.-à-d. qu'elle représente une information quantitative propre quoique cachée, information susceptible d'enrichir l'information globale de la série. Les exemples présentés dans cet essai illustrent cette condition : données manquantes dans une zone définie extrême de la variable, voire manquantes dans une zone interne doublement délimitée, données manquantes par interruption d'une série structurée à valeurs systématiquement croissantes ou décroissantes, etc. La seconde condition repose sur l'existence et le recours à un modèle distributionnel approprié, que ce modèle soit vérifié ou seulement stipulé. Grâce aux propriétés du modèle, notamment celles appliquées aux zones lacunaires concernées, la valeur moyenne probable des données manquantes peut être considérée, et surtout la valeur du ou des paramètres du modèle lui-même peut être estimée, palliant ainsi aux trois autres solutions qui confrontent le chercheur : renoncer à ses données et recommencer, baser ses estimations sur une série de données incomplète et appauvrie, procéder à une estimation structurellement biaisée. L'estimation par maximum de vraisemblance, une approche spécifiquement basée sur un modèle de probabilité, reste donc une bonne solution à considérer.

Dans cet essai, nous avons exploré différentes applications de l'estimation par maximum de vraisemblance en nous référant à autant de modèles de probabilité, la plupart basés sur le processus de Bernoulli. Ce choix nous a semblé justifié par notre expérience, les chercheurs se trouvant souvent démunis lorsque confrontés à des séries statistiques complètes ou incomplètes issues de ce processus : la tentation est forte d'exploiter directement la variable de surface, le "x" ou le "n", dont les propriétés statistiques sont souvent presque toujours incompatibles avec les procédures de tests usuelles (notamment en raison d'une forte asymétrie) et dont l'interprétation est délicate. D'un autre côté, l'estimation du paramètre responsable sous-jacent, l'espérance dans certains cas ou, dans les cas Bernoulli, la probabilité π, fournit une valeur à interprétation directe et aux propriétés distributionnelles plus malléables pour les traitements ultérieurs. Bien sûr, d'autres modèles que ceux présentés existent, et d'autres contextes d'estimation que ceux illustrés. Nous espérons néanmoins que les procédures et les exemples détaillés ici permettront aux collègues chercheurs de se dépanner dans les situations problématiques qu'ils sont susceptibles de rencontrer ou, à tout le moins, qu'ils leur serviront de tremplin pour trouver les solutions aux situations trop nombreuses que nous n'avons pu aborder.

References

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs (NJ): Prentice-Hall.

Burrows, P. M. (1972). Expected selection differentials for directional selection. *Biometrics*, *28*, 1091–1100.

Burrows, P. M. (1975). Variances of selection differentials in normal samples. *Biometrics*, *31*, 125–133.

de Moivre, A. (1758). The doctrine of chances.

Dixon, W. J. (1960). Simplified estimation from censored normal samples. *Annals of Mathematical Statistics*, *31*(385-391), 1895–1951.

[13] Une forme de winsorisation consiste, pour une variable continue, à remplacer une donnée occulte située à une extrémité de la série par la valeur extrême la plus proche dans la série. Dans le cas présent concernant une variable discrète (un nombre de pannes), la plus petite valeur maximale possible est évidemment 9+1.

Feller, W. (1968). *An introduction to probability theory and its applications, vol* (3e). I New York: Wiley.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions*. Wiley.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*. Wiley.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate discrete distributions (2e edition)*. New York: Wiley.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics*. Charles Grifin and Co Ltd.

Koutras, M. V. (1996). On a waiting time distribution in a sequence of bernoulli trials. *Annals of the Institute of Statistical Mathematics*, *48*, 789–806.

Laurencelle, L. (1983). La loi de pascal restreinte. *Lettres Statistiques*, *7*, 1–22.

Laurencelle, L. (1998). La variable n(k, 2) de la loi de pascal restreinte, avec compléments. *Lettres Statistiques*, *10*, 67–84.

Laurencelle, L. (2012). La loi de pascal restreinte et ses cas particuliers. *Tutorials in Quantitative Methods for Psychology*, *8*, 35–51.

Laurencelle, L. (2016). *L'étalonnage et la décision psychométrique (2e édition)*. Québec: Presses de l'Université du Québec.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley: New York.

Mood, A. M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics*, *11*, 367–392.

Philippou, A. N., Georghiu, C., & Philippou, G. N. (1983). A generalized geometric distribution and some of its properties. *Statistics and Probability Letters*, *1*, 171–175.

Rohatgi, V. K. (1976). *An introduction to probability and mathematical statistics*. New York: Wiley.

Uspensky, J. V. (1937). *Introduction to mathematical probability*. New York: McGraw-Hill.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

Laurencelle, L. (2019). Suppléer aux données manquantes par maximum de vraisemblance : Une application à des variables de Bernoulli//Making up for missing data by maximum likelihood estimation: An application to Bernoulli variables. *The Quantitative Methods for Psychology*, *15*(3), 188–199. doi:10.20982/tqmp.15.3.p168

Received: 26/04/2019 ~ Accepted: 23/09/2019