Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Estimation de la somme de variables aléatoires indépendantes et de même loi : comparaison de trois méthodes sur un exemple réel

par Miriam Basso et Bernard Beauzamy
Société de Calcul Mathématique S. A.

décembre 2008

en application du contrat Réseau Ferré de France
08 ETG000586 8004533, notifié le 21 juillet 2008

Résumé

Dans le cadre d'un travail réalisé pour RFF (Réseau Ferré de France), nous avons étudié le nombre de trains touchés lorsqu'un incident se produit ; un incident peut concerner plusieurs trains, le retard du premier ayant une conséquence sur les suivants. Ce nombre est une variable aléatoire X, à valeurs entières.

Si on s'intéresse aux conséquences d'un nombre fixé N d'incidents, le nombre de trains touchés correspondant est la somme $S_N = \sum_{i=1}^{N} X_i$. RFF veut disposer d'un encadrement pour cette somme, par exemple avec une probabilité de 95 %. On admet ici que, dans cette somme, les variables X_i sont indépendantes et de même loi que X. Cette loi est fournie par le retour d'expérience de RFF.

Souvent, on considère que la somme S_N suit une loi de Gauss : l'encadrement cherché s'en déduit immédiatement. Cette simplification est fausse. Le théorème central limite assure que le quotient $\frac{S_N}{N}$ converge en loi vers une gaussienne, mais non la variable S_N elle-même.

Dans le présent travail, nous fixons $N = 100$ et nous déterminons un encadrement à 95 % pour la somme S_N, par trois méthodes :

- Construction de la vraie loi de probabilité de la somme (par récurrence) ;
- Simulation de la somme par la méthode de Monte-Carlo (répétition de tirages indépendants selon la loi de X et somme de ces tirages) ;
- En faisant l'hypothèse que la somme suit une loi normale.

La simulation de Monte-Carlo converge bien vers la vraie loi de probabilité ; cette convergence est très lente (elle demande un nombre de tirages important), mais visible.

En faisant l'hypothèse que la somme de 100 répétitions de X suit une loi normale, nous obtenons une loi de probabilité différente de la vraie loi. L'encadrement obtenu est décalé vers la gauche ; la loi de la somme apparaît comme symétrique, alors qu'elle ne l'est pas en réalité

Voici l'encadrement de S_{100} à 95% obtenu par les trois méthodes :

	Borne inférieure	Borne supérieure	Ecart entre les bornes
Méthode 1	173	418	245
Méthode 2	171	417	246
Méthode 3	124	368	244

Encadrement de S_{100} à 95 %, comparaison des trois méthodes

Des estimations fines, comme par exemple celle du dépassement d'un seuil, $P\{S_N > a\}$ ($a > 0$ quelconque), ne seront pas correctes sous l'hypothèse gaussienne. Par exemple, $P\{S_{100} > 246\}$ est estimée à 51 % sous cette hypothèse, alors que cette probabilité est en fait égale à 39 %.

I. Position du problème

On note X la variable aléatoire "nombre de trains touchés lors d'un incident" ; elle est à valeurs entières x_k, $1 \le x_k \le 237$. Nous notons p_k la probabilité de x_k. Les valeurs des p_k sont données en annexe ; elles proviennent du retour d'expérience de RFF. Voici l'histogramme :

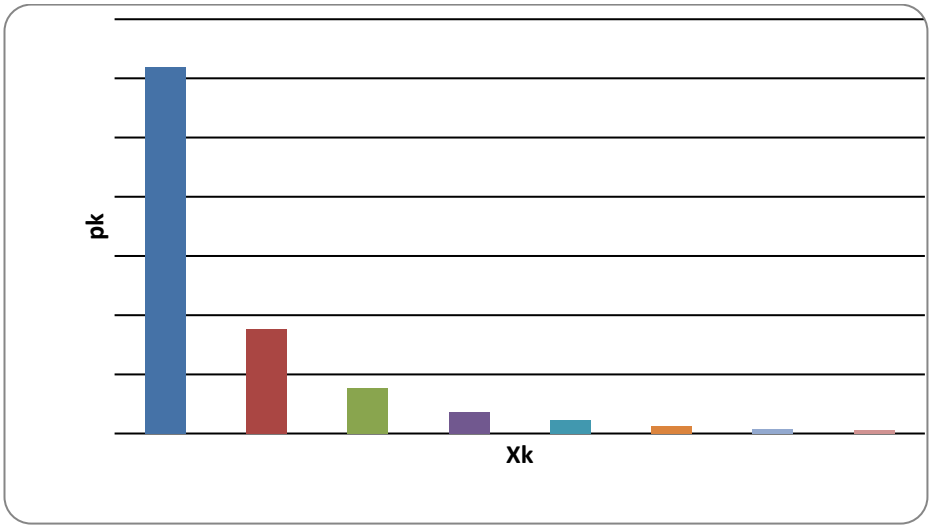

Figure 1 : Loi de probabilité de X

Cette variable admet pour moyenne $m = 2$ et pour écart-type $\sigma = 6$. Nous voulons déterminer la loi de la variable aléatoire S_{100} définie par :

$$S_{100} = \sum_{i=1}^{100} X_i .$$

L'ensemble des valeurs possibles pour la variable aléatoire S_{100} est l'intervalle $100 - 2370$. Nous recherchons un intervalle de confiance à 95 %.

II. Méthodes testées

A. *Méthode 1: construction de la loi de probabilité de S_N*

Une première méthode consiste à déterminer de façon exacte la loi de probabilité de S_{100} **:** nous calculons par récurrence la probabilité des sommes S_N pour $N = 1,...,100$.

Initialisation

Les valeurs de S_1 sont les entiers compris dans l'intervalle $1 - 237$. Pour $k = 1,...,237$, $P\{S_1 = k\} = p_k$.

Relation de récurrence

Pour $N = 2,...,100$, nous déterminons l'ensemble des valeurs de S_N et leurs probabilités à partir de S_{N-1}. On a $S_N = j$, où j est dans l'intervalle $[N, 237N]$ si $S_{N-1} = j - k$ et $X_N = k$, donc :

$$P(S_N = j) = \sum_{k=A}^{B} P(S_{N-1} = j-k) p_k \qquad (1)$$

avec :

$$A = max\{1, j-(N-1)M\}$$
$$B = min\{M, j-N+1\}$$

où $M = 237$.

La formule de récurrence (1) permet de déterminer la loi de probabilité de S_N pour tout N.

Pour obtenir un encadrement de S_{100} à 95 % nous construisons le cumul des probabilités et déterminons les deux quantiles correspondant à une probabilité de 97,5 % et 2,5 %.

B. Méthode 2: simulation de Monte-Carlo

Simulation de la variable X

Pour simuler la variable aléatoire X par la méthode de Monte-Carlo, nous tirons de façon aléatoire un nombre β compris entre 0 et 1. La valeur x_k résultant de cette simulation correspond à la valeur associée à la probabilité p_k, où k est l'entier vérifiant :

$$\sum_{j=1}^{k-1} p_j \le \beta < \sum_{j=1}^{k} p_j$$

En répétant ce tirage M fois, nous construisons une loi de probabilité "simulée" pour X ; elle converge vers la vraie loi quand $M \to \infty$.

Dans l'exemple étudié, il faut répéter la simulation 10^8 fois, pour que l'écart entre les probabilités réelles et les probabilités obtenues par la simulation soit de l'ordre de 1 % pour tous les p_k. La convergence est donc très lente.

Simulation de la somme

Une simulation de S_{100} correspond à la somme de 100 valeurs X_k obtenues par une simulation de X. En répétant ceci 10^6 fois, nous construisons la loi de probabilité de S_{100}. L'encadrement de S_{100} est obtenu en faisant le cumul des probabilités et en cherchant les deux quantiles correspondant à 97,5 % et 2,5 % de probabilité cumulée.

C. Méthode 3: loi de Gauss

Soit X une variable aléatoire de moyenne m et d'écart-type σ. D'après le théorème central limite, pour N suffisamment grand, la variable :

$$Z_N = \frac{\frac{1}{N}\left(\sum_{i=1}^{N} X_i\right) - m}{\sigma / \sqrt{N}} \tag{2}$$

suit approximativement une loi gaussienne centrée réduite. Il en résulte que :

$$P\{a \le Z_N \le b\} \approx \int_a^b h(t)dt \tag{3}$$

où h est la densité de la loi de Gauss, $h(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$.

Le sens précis de l'énoncé (3) est que l'égalité est de mieux en mieux vérifiée lorsque $N \to +\infty$, si a et b sont fixés, indépendants de N. Si on considère abusivement (3) comme une égalité, on peut écrire :

$$P\left\{m\,N + a\,\sigma\sqrt{N} \le S_N \le m\,N + b\,\sigma\sqrt{N}\right\} = \int_a^b h(t)dt$$

ou encore, avec :

$$\alpha = m\,N + a\,\sigma\sqrt{N}\,,\ \beta = m\,N + b\,\sigma\sqrt{N}\,, \tag{4}$$

$$P\{\alpha \le S_N \le \beta\} = \int_{\frac{\alpha - m\,N}{\sigma\sqrt{N}}}^{\frac{\beta - m\,N}{\sigma\sqrt{N}}} h(t)\,dt$$

ou encore :

$$P\{\alpha \le S_N \le \beta\} = \int_\alpha^\beta h_{m\,N,\sigma\sqrt{N}}(t)dt$$

où :

$$h_{m,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(t-m)^2}{2\sigma^2}\right).$$

Cela revient à dire que S_N suit une loi de Gauss de paramètres $m\,N, \sigma\sqrt{N}$. Ceci est incorrect et l'erreur commise tient au fait que α et β ne devraient pas dépendre de N, alors qu'ils en dépendent dans (4).

Encadrer S_N avec une probabilité de 95 % revient à se placer dans l'intervalle : $\left[m\,N - 2\sigma\sqrt{N},\ m\,N + 2\sigma\sqrt{N} \right]$.

III. Résultats

A. Comparaison des lois de probabilité

A partir des trois méthodes présentées, nous avons déterminé la loi de probabilité de S_{100}. La moyenne et l'écart type obtenus sont donnés dans le tableau ci-dessous.

	Moyenne	Ecart-type
Méthode 1	246,33	61,98
Méthode 2	246,34	61,97
Méthode 3	246,33	61,98

Tableau 1 : Moyenne et écart type de S_{100}, comparaison des trois méthodes

Le graphique ci-dessous représente les lois de probabilités correspondantes.

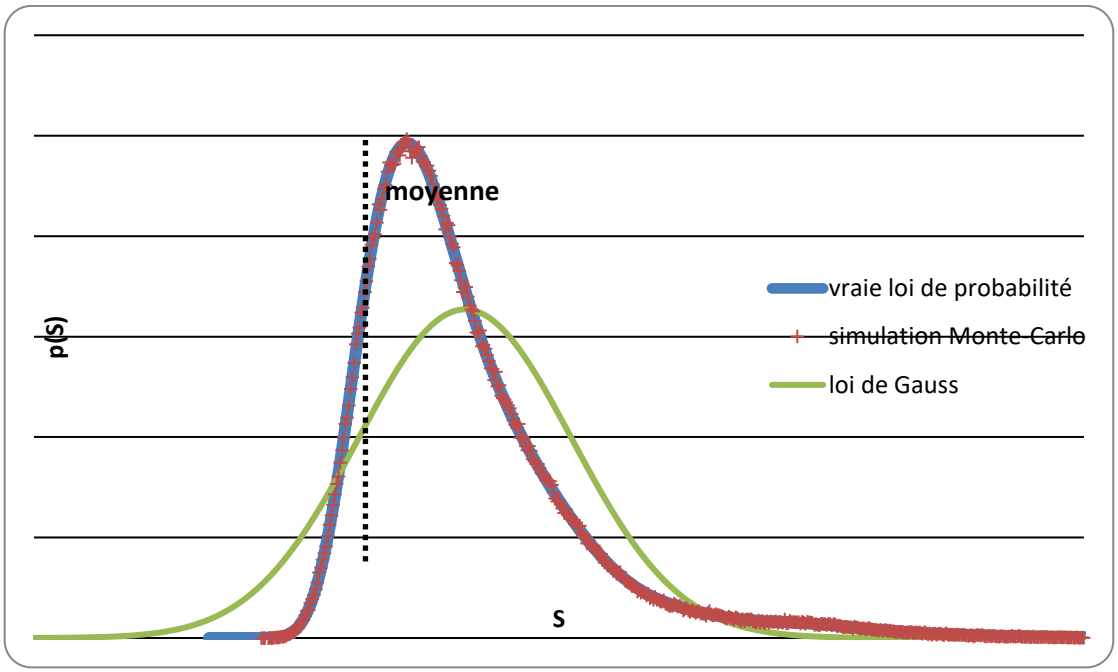

Figure 2 : Loi de probabilité de S_{100}, comparaison des trois méthodes

Bien que la valeur moyenne de S_{100} soit la même, la représentation graphique met en évidence la diversité des résultats obtenus par les trois méthodes. Alors que la loi de probabilité obtenue par la méthode 3 (gaussienne) est parfaitement symétrique et centrée sur la moyenne, les deux autres sont dissymétriques. Les courbes obtenues par les méthodes 1 et 2 sont très proches : les deux courbes sont confondues quand le nombre de tirages aléatoires correspondant à la méthode de Monte-Carlo tend vers l'infini.

B. Comparaison de l'encadrement à 95 %

L'encadrement de S_{100} à 95% obtenu par les trois méthodes est présenté ci-dessous.

	Borne inférieure	Borne supérieure	Ecart entre les bornes
Méthode 1	173	418	245
Méthode 2	171	417	246
Méthode 3	124	368	244

Tableau 2 : Encadrement de S_{100} à 95 %, comparaison des trois méthodes

Les trois intervalles obtenus se recoupent et ont une largeur du même ordre de grandeur. L'encadrement de S_{100} obtenu par la simulation de Monte-Carlo est proche de l'encadrement obtenu à partir de la vraie loi de probabilité. L'encadrement obtenu par la méthode 3, à savoir $124-368$, est décalé par rapport aux deux autres.

Annexe

Voici les valeurs des probabilités des différentes valeurs prises par X.

x_k	p_k
1	0.6195281
2	0.17665321
3	0.07668426
4	0.03678982
5	0.02297423
6	0.01334989
7	0.00776156
8	0.00651971
9	0.00667495
10	0.00512263

x_k	p_k
27	0.00046569
29	0.00015523
30	0.00031046
31	0.00031046
32	0.00015523
33	0.00015523
34	0.00031046
35	0.00015523
36	0.00015523
39	0.00031046

11	0.00434648
12	0.0024837
13	0.00357032
14	0.00170754
15	0.00093139
16	0.00093139
17	0.00077616
18	0.00046569
19	0.00155231
20	0.00093139
21	0.00155231
22	0.00077616
23	0.00062093
24	0.00046569
25	0.00062093
26	0.00031046
27	0.00046569

40	0.00015523
43	0.00031046
45	0.00015523
48	0.00015523
50	0.00015523
58	0.00015523
60	0.00015523
63	0.00031046
69	0.00031046
76	0.00015523
79	0.00015523
83	0.00031046
88	0.00031046
96	0.00015523
112	0.00015523
182	0.00015523
237	0.00015523

Histogramme réel de X