Sur l'approximation de la loi binomiale par la loi normale.

Auteur : **Christian Maillard**
840 B, avenue des Serrets - 04100 Manosque
(trueil@wanadoo.fr)

Introduction

Nous citons dans cette introduction une note de Paul-Louis HENNEQUIN parue dans le bulletin vert n°436 de Novembre-Décembre 2001 (p.732-733), suite à l'article de Louis-Marie BONNEVAL intitulé "Intervalles : de confiance ?" et référencé ci-dessous.
Nous invitons vivement le lecteur à prendre connaissance de ce texte de L. M. Bonneval, avant de poursuivre la lecture du présent article.

Dans le bulletin vert n°427 (mars-avril 2000), pages 141-170, Louis-Marie BONNEVAL s'intéresse à l'intervalle de confiance pour l'estimation de la probabilité p inconnue d'un événement aléatoire à partir de la réalisation de n expériences indépendantes dans lesquelles l'événement est apparu avec une fréquence f.
Il propose les deux conjectures suivantes :

1)- Pour tout naturel $n > 0$, et pour tout réel p de [0,1] :

$$P\left\{ \left| f - p \right| \leq \frac{1}{\sqrt{n}} \right\} > 0.9$$

résultat facile à mémoriser pour un élève débutant en statistique inférentielle.

2)- Pour tout naturel $n > 20$, et pour tout réel p de [0,1] :

$$P\left\{ \left| f - p \right| \leq \frac{1}{\sqrt{n}} \right\} > 0.93$$

et suggère de travailler sur le niveau 0.93.

Daniel SAADA (Saada@club-internet.fr) nous a adressé le 17 Février 2001 une étude relative au cas particulier $p = 0.5$ et $n = 4\,k^2$, k naturel. Il obtient alors :

$$P\left\{ \left| f - p \right| \leq \frac{1}{\sqrt{n}} \right\} > 0.915$$

Par ailleur, il montre, toujours dans le cas particulier $p = 0.5$ que, pour $n > 5$:

$$P\left\{ \left| f - p \right| \leq \frac{1}{\sqrt{n}} \right\} > 0.929$$

De son côté, Christian MAILLARD nous a adressé le 6 Août 2001 la démonstration des inégalités ci-dessous qui incluent les conjectures de L. M. Bonneval :

Pour tout p réel de [0,1] et pour tout naturel n supérieur à 552,

$$P\left\{ \left| f - p \right| \leq \frac{1}{\sqrt{n}} \right\} > 0.95$$

Pour tout p réel de [0,1] et pour tout naturel n supérieur à 56,

$$P\left\{ \left| f - p \right| \leq \frac{1}{\sqrt{n}} \right\} > 0.94$$

Pour tout p réel de [0,1] et pour tout naturel n supérieur à 20,

$$P\left\{ \left| f-p \right| \le \frac{1}{\sqrt{n}} \right\} > 0.93$$

Et enfin pour tout p réel de [0,1] et pour tout naturel n non nul,

$$P\left\{ \left| f-p \right| \le \frac{1}{\sqrt{n}} \right\} > 0.9$$

Cette démonstration s'accompagne d'une analyse très précise de l'approximation de la loi binomiale par la loi de Laplace-Gauss.

La démonstration de la conjecture Bonneval s'appuie sur le principe (mais s'appuie seulement) du calcul de l'erreur commise en remplaçant la loi binomiale par la loi normale.
La partie qui suit donne un calcul général de l'erreur entre les deux lois (une majoration de l'erreur, bien entendu).
Pour aboutir ensuite à la conjecture Bonneval, il faudra choisir $b = \frac{1}{\sqrt{pq}}$ mais aussi faire un calcul spécifique de l'erreur dans ce cas précis (qui ne sera pas développé ici).

Calcul de l'erreur sur Prob($|\xi_n^*| \le b$) où b est un réel positif quelconque

S_n suit une loi binomiale de paramètres n et p, et on pose $\xi_n^* = \frac{S_n - np}{\sqrt{npq}}$

Il est impossible de développer complètement les calculs menant à cet encadrement de l'erreur.
Je voudrais indiquer l'idée de départ et donner les résultats que j'ai trouvés.

Préalable : Les conditions sur n et p seront les suivantes : $npq \ge \max\left(10\, b^2 \,;\, \frac{b^2}{25} \,;\, 116 \,;\, \frac{4}{b^2}\right)$

$$A = p\left(\left| \xi_n^* \right| \le b \right) = \sum_{\mathrm{np}-b\sqrt{\mathrm{npq}} \le k \le \mathrm{np}+b\sqrt{\mathrm{npq}}} C_n^k\, p^k\, q^{n-k} = \sum_{\mathrm{np}-b\sqrt{\mathrm{npq}} \le k \le \mathrm{np}+b\sqrt{\mathrm{npq}}} \frac{n!}{k!\,(n-k)!}\, p^k\, q^{n-k}$$

On utilise alors l'encadrement suivant de $n!$ (c.f. par exemple "Méthodes Stochastiques" de ZieZold-KriKenberg)

$$n^{n+0.5}\, e^{-n}\, e^{\frac{1}{12n+1}}\, \sqrt{2\pi} \le n! \le n^{n+0.5}\, e^{-n}\, e^{\frac{1}{12n}}\, \sqrt{2\pi}$$

Dans ce cas :

$$A \ge \frac{e^{\frac{1}{12n+1}}}{\sqrt{2\pi \mathrm{npq}}} \sum_{\mathrm{np}-b\sqrt{\mathrm{npq}} \le k \le \mathrm{np}+b\sqrt{\mathrm{npq}}} \left(\frac{\mathrm{np}}{k}\right)^{k+0.5} \left(\frac{q}{1-\frac{k}{n}}\right)^{n-k+0.5} e^{-\frac{1}{12k}-\frac{1}{12(n-k)}}$$

qui donne, après une étude de la dernière exponentielle :

$$A \ge \frac{\alpha_n}{\sqrt{2\pi \mathrm{npq}}} \sum_{\mathrm{np}-b\sqrt{\mathrm{npq}} \le k \le \mathrm{np}+b\sqrt{\mathrm{npq}}} \left(\frac{\mathrm{np}}{k}\right)^{k+0.5} \left(\frac{q}{1-\frac{k}{n}}\right)^{n-k+0.5}$$

$$\text{avec } \alpha_n = \exp\left(\frac{1}{12n+1} - \frac{1}{12n\left(\mathrm{pq} - (q-p)\, b \sqrt{\frac{\mathrm{pq}}{n} - \frac{b^2\,\mathrm{pq}}{n}}\right)} \right)$$

Cette écriture n'étant valable que si $\mathrm{np} - b\sqrt{\mathrm{npq}} > 0$ (1) et $\mathrm{np} + b\sqrt{\mathrm{npq}} < n$ (2).
Les conditions (1) et (2) sont vérifiées.

Je vais utiliser la méthode du point médian. On a :

$$\sum_{np-b\sqrt{npq}\le k\le np+b\sqrt{npq}} \left(\frac{np}{k}\right)^{k+0.5}\left(\frac{q}{1-\frac{k}{n}}\right)^{n-k+0.5} = n\sum_{np-b\sqrt{npq}\le k\le np+b\sqrt{npq}} \frac{1}{n} f\left(\frac{k}{n}\right)$$

$$\text{avec } f(t)=\left(\frac{p}{t}\right)^{nt+0.5}\left(\frac{q}{1-t}\right)^{n-nt+0.5}$$

On a alors $\left|\int_{\frac{k-0.5}{n}}^{\frac{k+0.5}{n}} f(t)\,dt-\frac{1}{n}f\left(\frac{k}{n}\right)\right| \le \frac{K}{24n^3}$ où K est un majorant de $|f''(t)|$ et $p-b\sqrt{\frac{pq}{n}} \le p+b\sqrt{\frac{pq}{n}}$. K dépend de n, p et b.

On fait alors varier k et pour cela je suis amené à distinguer deux cas : le cas où $np-b\sqrt{npq}$ est un entier, et le cas où c'en est pas un.
Je pose alors $\alpha=p-\frac{[np+b\sqrt{npq}]+0.5}{n}$ et $\beta=p-\frac{[np-b\sqrt{npq}]+0.5}{n}$ si $np-b\sqrt{npq}$ n'est pas entier, sinon $\beta=b\sqrt{\frac{pq}{n}}+\frac{0.5}{n}$.
Dans les deux cas, il est facile de voir que $-b\sqrt{\frac{pq}{n}}-\frac{0.5}{n}\le\alpha\le -b\sqrt{\frac{pq}{n}}+\frac{0.5}{n}$ et $b\sqrt{\frac{pq}{n}}-\frac{0.5}{n}\le\beta\le b\sqrt{\frac{pq}{n}}+\frac{0.5}{n}$.

On obtient alors $A\ge\alpha_n\sqrt{\frac{n}{2\pi}}\int_\alpha^\beta e^{K(p,v)}\,dv-\frac{\alpha_n K\mu}{24n^2}$ avec $e^{K(p,v)}=\left(\frac{p}{p-v}\right)^{np-nv}\left(\frac{q}{q+v}\right)^{nq+nv}\frac{1}{\sqrt{(p-v)(q+v)}}$ et $\mu=\frac{2b\sqrt{npq}+1}{\sqrt{2\pi npq}}$.

On développe *K(p,v)* sous la forme : $K(p,v)=-\frac{1}{2}\ln pq-\frac{nv^2}{2pq}+X$ où X dépend de n, v, p, q.

Alors $A\ge\alpha_n\sqrt{\frac{n}{2\pi}}\int_\alpha^\beta e^{K(p,v)}\,dv-\frac{\alpha_n K\mu}{24n^2}\ge\alpha_n\sqrt{\frac{n}{2\pi pq}}\int_\alpha^\beta e^{-\frac{nv^2}{2pq}}e^X\,dv-\frac{K\mu}{24n^2}$ (car α_n est inférieur à 1), et on obtient de même une majoration similaire $A\le\sqrt{\frac{n}{2\pi pq}}\int_\alpha^\beta e^{-\frac{nv^2}{2pq}}e^X\,dv+\frac{K\mu}{24n^2}$.

Les conditions que j'ai imposées au départ sur n, p, q et b permettent d'encadrer les termes X et K et alors de calculer l'erreur commise en remplaçant la loi binomiale par la loi normale.
Voici mes résultats :

$$\left|\mathbf{Prob}\left(|\xi_n^*|\le b\right)-\frac{1}{\sqrt{2\pi}}\int_{-b}^{b}e^{-\frac{u^2}{2}}\,du\right|\le\frac{0,1062+e^{-\frac{1}{2}\left(b-\frac{0,5}{\sqrt{npq}}\right)^2}}{\sqrt{2\pi npq}}\quad \textbf{pour } b\le 3$$

$$\left|\mathbf{Prob}\left(|\xi_n^*|\le b\right)-\frac{1}{\sqrt{2\pi}}\int_{-b}^{b}e^{-\frac{u^2}{2}}\,du\right|\le\frac{0,097}{\sqrt{npq}}\quad \textbf{pour } b\ge 3$$

Exemple : p=0,03 ; n=4000 ; np=120 ; npq=116,4 ; $b=\frac{10}{\sqrt{116,4}}$. On a :

$$A=\text{Prob}\left(110\le\sum_{i=0}^{4000}X_i\le 130\right)=\text{Prob}\left(\left|\xi_{4000}^*\right|\le\frac{10}{\sqrt{116,4}}\right)$$

le calcul donne une erreur inférieure à 0,02902 et donc une valeur de A comprise entre 0,6751 et 0,61699.
La valeur réelle est d'environ 0,66969. Par conséquent l'approximation est relativement précise compte tenu du caractère un peu "limite" car la valeur minimale d'utilisation est *npq* supérieur à 116, et ici on a 116,4.