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

EvalRisk_2021 V2 : présentation théorique

La situation est la suivante : on se dispose à mettre sur le marché un lot constitué de N_1 appareils et on se demande combien, parmi eux, ne vont pas fonctionner correctement. On dispose d'une information préalable, portant sur un lot (généralement beaucoup plus petit) de N appareils : on a constaté que sur ce lot préliminaire de taille N, une quantité n d'appareils n'avaient pas fonctionné correctement. Voir la présentation générale pour les situations concrètes qui peuvent se rencontrer.

1. La formule de définition

Si on connaissait la probabilité de panne, notée λ, pour chaque appareil, la réponse serait donnée par une loi binomiale : la probabilité que n_1 appareils tombent en panne est donnée par la formule :

$$P(n_1) = \binom{N_1}{n_1} \lambda^{n_1} (1-\lambda)^{N_1-n_1} \tag{1}$$

Ceci sous réserve que :

- On spécifie un intervalle de temps pour l'observation : généralement une année. Si on attend indéfiniment, tout appareil tombera en panne !
- Les pannes des appareils puissent être considérées comme indépendantes ;
- On ne se soucie d'aucune cause explicative relative aux pannes : il ne s'agit pas de savoir si certains appareils sont au soleil, d'autres exposés à la pluie, s'ils sont utilisés par des opérateurs compétents ou non, etc.

L'approche par loi binomiale est donc une approche résolument grossière, qui ne constate que les faits sans se soucier d'aucune explication. A ce titre, elle a vocation à être préliminaire. Du reste, en général, en ce qui concerne un lot futur, on ne dispose d'aucune information préalable : l'approche retenue ici est donc pertinente.

Mais, dans la situation présente, la valeur précise de λ est inconnue. Voyons progressivement comment incorporer l'information résultant du lot initial.

Bien entendu, dans les faits, λ a vocation à prendre toute valeur entre 0 et 1, mais imaginons, pour les besoins du calcul, que λ ne puisse prendre que deux valeurs, par exemple $\lambda_1 = \frac{1}{3}$ et $\lambda_2 = \frac{1}{2}$, avec probabilités respectives q_1, q_2, $q_1 + q_2 = 1$. La formule (1) serait remplacée par :

$$P(n_1) = \binom{N_1}{n_1} \left(\lambda_1^{n_1} (1-\lambda_1)^{N_1-n_1} q_1 + \lambda_2^{n_1} (1-\lambda_2)^{N_1-n_1} q_2 \right) \tag{2}$$

En effet, à chacune des deux situations correspond une valeur de $P(n_1)$ et chaque situation doit être pondérée par la probabilité correspondante.

Plus généralement, si λ pouvait prendre K valeurs, $\lambda_1, ..., \lambda_K$, avec probabilités respectives $q_1, ..., q_K$, la formule (2) serait remplacée par :

$$P(n_1) = \binom{N_1}{n_1} \sum_{k=1}^{K} \lambda_k^{n_1} (1-\lambda_k)^{N_1-n_1} q_k \tag{3}$$

A partir de la formule (3), le passage du discret au continu se fait facilement. Admettons que λ puisse prendre toute valeur entre 0 et 1, avec une densité de probabilité $f(\lambda)$; la formule (3) devient :

$$P(n_1) = \binom{N_1}{n_1} \int_0^1 \lambda^{n_1} (1-\lambda)^{N_1-n_1} f(\lambda) d\lambda \tag{4}$$

La formule (4) nous permet d'incorporer l'information préliminaire : n défauts sur le lot initial de taille N. En effet, nous savons que, dans ces conditions, la densité de probabilité de λ est donnée par la fonction :

$$f_{n,N}(\lambda) = c\lambda^n (1-\lambda)^{N-n}$$

où c est une constante de normalisation :

$$c = \frac{(N+1)!}{n!(N-n)!}$$

Voir le livre [MPPR] pour le détail des calculs.

Reportant dans (4), nous obtenons la formule :

$$P(n_1) = c\binom{N_1}{n_1}\int_0^1 \lambda^{n_1}(1-\lambda)^{N_1-n_1}\,\lambda^{n}(1-\lambda)^{N-n}\,d\lambda = c\binom{N_1}{n_1}\int_0^1 \lambda^{n+n_1}(1-\lambda)^{N-n+N_1-n_1}\,d\lambda$$

Le calcul de l'intégrale se fait par récurrence : voir [MPPR]. On obtient finalement une formule explicite :

$$P(n_1) = \frac{N+1}{N+N_1+1}\frac{\binom{N_1}{n_1}\binom{N}{n}}{\binom{N+N_1}{n+n_1}} \tag{5}$$

2. Comment calculer à partir de cette formule ?

En pratique, cette formule ne pourra pas être utilisée directement : si N_1 est grand (par exemple de l'ordre du million), les coefficients du binôme $\binom{N_1}{n_1}$ et $\binom{N+N_1}{n+n_1}$ sont des nombres énormes, qui ne peuvent être traités directement par un ordinateur. Il est donc nécessaire de rechercher des formes plus utilisables de la formule (5).

Sous forme de factorielles, la formule (5) s'écrit :

$$\begin{aligned} P(n_1) &= \frac{N+1}{N+N_1+1}\frac{N_1!}{(N_1-n_1)!n_1!}\frac{N!}{(N-n)!n!}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N+N_1)!} \\ &= \frac{N_1!}{(N_1-n_1)!n_1!}\frac{(N+1)!}{(N-n)!n!}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N+N_1+1)!} \end{aligned}$$

et sous forme de produit :

$$P(n_1) = \frac{\prod_{j=1}^{n_1}(N_1-n_1+j)\prod_{j=1}^{n+1}(N-n+j)\prod_{j=1}^{n_1}(n+j)}{\prod_{j=1}^{n+n_1+1}(N+N_1-n-n_1+j)\prod_{j=1}^{n_1} j}$$

C'est déjà beaucoup plus simple, mais reste inutilisable en pratique : si $N_1 = 10^6$ et si $n_1 = 10^3$, un produit du type $\prod_{j=1}^{n_1}(N_1-n_1+j)$ vaudra environ 10^{6000}.

On peut vouloir faire un calcul par récurrence ; on obtient immédiatement :

$$\frac{P(n_1+1)}{P(n_1)} = \frac{N_1-n_1}{N+N_1-n-n_1}\frac{n+n_1+1}{n_1+1} \tag{6}$$

formule très simple, qui devrait permettre de calculer chaque $P(n_1)$ à partir de $P(0)$:

$$P(0)=\frac{(N+1)!}{(N-n)!}\frac{(N+N_1-n)!}{(N+N_1+1)!}=\prod_{j=1}^{n+1}\frac{N-n+j}{N+N_1-n+j} \tag{7}$$

Malheureusement, comme expliqué plus haut, $P(0)$ est extrêmement petit ; il est considéré comme nul par l'ordinateur, qui en déduit que tous les $P(n_1)$ sont nuls : la formule de récurrence est inutilisable en pratique sous forme directe : il faut calculer sur les logarithmes.

On commence par $Log\left(P(0)\right)$ à partir de la formule (7) :

$$Log\left(P(0)\right)=\sum_{j=1}^{n+1} Log(N-n+j)-Log(N+N_1-n+j)$$

Ensuite, on calcule chaque $Log\left(P(n_1)\right)$ à partir du précédent, en utilisant la formule (6) :

$$Log\left(P(n_1+1)\right)=Log\left(P(n_1)\right)+Log(N_1-n_1)+Log(n+n_1+1)-Log(N+N_1-n-n_1)-Log(n_1+1)$$

On en déduit la valeur de $P(n_1)$ en prenant l'exponentielle et la valeur de la fonction de répartition, en sommant à partir de 0. Aucune approximation n'est faite, et le calcul a été testé jusqu'à $N_1=10^9$; il prend moins de 30 secondes.

3. Maximum de probabilité

On peut se demander pour quelle valeur de n_1 la probabilité $P(n_1)$ est maximale. La réponse est facile ; la condition $P(n_1+1)>P(n_1),$ en utilisant (6), équivaut à :

$$(N_1-n_1)(n+n_1+1)>(N+N_1-n-n_1)(n_1+1)$$

c'est-à-dire :

$$(N_1+1)n>N(n_1+1),\ n_1+1<(N_1+1)\frac{n}{N}$$

La valeur maximale est donc obtenue approximativement pour $n_1\approx N_1\frac{n}{N}$. Ceci est conforme à l'intuition : si 10 objets ont été défectueux sur 100, on s'attend à en avoir à peu près $\frac{10}{100}\times 2000=200$ sur 2000 : c'est une règle de trois.

4. Calcul de l'espérance de la loi

On part de l'expression, déduite de (1) :

$$P(n_1) = \frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!n_1!}$$

Puisqu'il s'agit d'une loi de probabilité, on a $\sum_{n_1=0}^{N_1} P(n_1) = 1,$ ce qui se traduit par :

$$\sum_{n_1=0}^{N_1} \frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!n_1!} = \frac{(N-n)!n!(N+N_1+1)!}{(N+1)!N_1!} \qquad (7)$$

L'espérance de la loi est donnée par la formule :

$$E = \sum_{n_1=0}^{N_1} n_1 P(n_1) = \frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}\sum_{n_1=1}^{N_1} \frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-1)!}$$

ce que l'on peut écrire :

$$E = \frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}\sum_{n_1=1}^{N_1} \frac{\big((N+1)+(N_1-1)-(n+1)-(n_1-1)\big)!\big((n+1)+(n_1-1)\big)!}{\big((N_1-1)-(n_1-1)\big)!(n_1-1)!}$$

avec $N' = N+1,\ N_1' = N_1-1,\ n' = n+1,\ n_1' = n_1-1,$ on obtient :

$$\sum_{n_1=1}^{N_1} \frac{\big((N+1)+(N_1-1)-(n+1)-(n_1-1)\big)!\big((n+1)+(n_1-1)\big)!}{\big((N_1-1)-(n_1-1)\big)!(n_1-1)!} = \sum_{n_1'=0}^{N_1'} \frac{(N'+N_1'-n'-n_1')!(n'+n_1')!}{(N_1'-n_1')!n_1'!}$$

et, en utilisant (7):

$$\sum_{n_1'=0}^{N_1'} \frac{(N'+N_1'-n'-n_1')!(n'+n_1')!}{(N_1'-n_1')!n_1'!} = \frac{(N'-n')!n'!(N'+N_1'+1)!}{(N'+1)!N_1'!} = \frac{(N-n)!(n+1)!(N+N_1+1)!}{(N+2)!(N_1-1)!}$$

ce qui donne :

$$E = \frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}\frac{(N-n)!(n+1)!(N+N_1+1)!}{(N+2)!(N_1-1)!}$$

et donc finalement :

$$E = \frac{(n+1)N_1}{N+2}$$

La différence entre le n_1 qui donne la plus grande probabilité (§3) et l'espérance est que le premier est nécessairement un entier, tandis que l'espérance est un nombre réel.

Si l'on pose $n=\alpha N,$ l'espérance s'écrit : $E=\dfrac{(\alpha N+1)N_1}{N+2}\approx \alpha N_1$ si N est grand. Approximativement, l'espérance ne dépend pas de n,N séparément, mais du quotient $\alpha=\dfrac{n}{N}$ et de N_1 (proportionnalité à N_1).

5. Calcul de la variance de la loi

On procède comme précédemment. On écrit :

$$E_2=\sum_{n_1=0}^{N_1} n_1^2 p(n_1)=\frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}\sum_{n_1=1}^{N_1} n_1\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-1)!}$$

D'où :

$$\sum_{n_1=1}^{N_1} n_1\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-1)!}=\sum_{n_1=1}^{N_1}(n_1-1+1)\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-1)!}$$
$$=\sum_{n_1=2}^{N_1}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-2)!}+\sum_{n_1=1}^{N_1}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-1)!}$$

On sait estimer le second terme :

$$B=\sum_{n_1=1}^{N_1}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-1)!}=\frac{(N-n)!(n+1)!(N+N_1+1)!}{(N+2)!(N_1-1)!}$$

Pour le premier, on écrit :

$$A=\sum_{n_1=2}^{N_1}\frac{(N+N_1-n-n_1)!(n+n_1)!}{(N_1-n_1)!(n_1-2)!}$$
$$=\sum_{n_1=2}^{N_1}\frac{((N+2)+(N_1-2)-(n+2)-(n_1-2))!((n+2)+(n_1-2))!}{((N_1-2)-(n_1-2))!(n_1-2)!}$$

On pose $N''=N+2,\ N_1''=N_1-2$, $n''=n+2,\ n_1''=n_1-2$; on a :

$$A=\sum_{n_1''=0}^{N_1''}\frac{\left(N''+N_1''-n''-n_1''\right)!\left(n''+n_1''\right)!}{\left(N_1''-n_1''\right)!n_1''!}=\frac{(N''-n'')!n''!(N''+N_1''+1)!}{(N''+1)!N_1''!}$$
$$=\frac{(N-n)!(n+2)!(N+N_1+1)!}{(N+3)!(N_1-2)!}$$

On obtient :

$$E_2=\frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}(A+B)$$
$$=\frac{(N+1)!}{(N-n)!n!}\frac{N_1!}{(N+N_1+1)!}\left(\frac{(N-n)!(n+2)!(N+N_1+1)!}{(N+3)!(N_1-2)!}+\frac{(N-n)!(n+1)!(N+N_1+1)!}{(N+2)!(N_1-1)!}\right)$$
$$=\frac{N_1(N_1-1)}{(N+2)}\frac{(n+1)(n+2)}{(N+3)}+\frac{N_1}{(N+2)}\frac{(n+1)}{(N+2)}$$
$$=\frac{(n+1)}{(N+2)}N_1\left(1+(N_1-1)\frac{(n+2)}{(N+3)}\right)$$

La variance est :

$$V=E_2-E^2=\frac{(n+1)}{(N+2)}N_1\left(1+(N_1-1)\frac{(n+2)}{(N+3)}\right)-\left(\frac{(n+1)N_1}{N+2}\right)^2$$

On obtient la formule exacte :

$$V=\frac{(n+1)N_1(N+N_1+2)(N-n+1)}{(N+2)^2(N+3)}$$

Pour N et N_1 grands, on a la formule approchée :

$$V\approx\frac{(n+1)N_1(N+N_1)(N-n)}{N^3}$$

Si on pose $n=\alpha N$,

$$V\approx\frac{(\alpha N+1)N_1(N+N_1)(N-\alpha N)}{N^3}\approx\frac{\alpha(1-\alpha)(N+N_1)N_1}{N}=\alpha(1-\alpha)\left(1+\frac{N_1}{N}\right)N_1$$

Cette quantité est maximale pour $\alpha=\frac{1}{2}$ (cas $n=\frac{N}{2}$), décroît lorsque N augmente et augmente comme N_1^2.

Si N_1 est grand devant N (ce qui est le cas en général), on obtient l'expression simplifiée :

$$V \approx \alpha(1-\alpha)\frac{N_1^2}{N}$$

Cette expression peut laisser croire que la variance est nulle si $\alpha = 0$ ou 1. Ce n'est pas le cas ; en revenant à l'expression complète, on a, si $\alpha = 0$, et donc $n = 0$:

$$V = \frac{(N+1)N_1(N+N_1+2)}{(N+2)^2(N+3)} \approx \frac{N_1(N+N_1)}{N^2}$$

et si $n = N$:

$V = \frac{(N+1)N_1(N+N_1+2)}{(N+2)^2(N+3)}$: l'expression est la même.

Si N et N_1 sont fixés, la variance est une fonction du second degré par rapport à n, concave, maximale pour $n = \frac{N}{2}$; voici un exemple pour $N = 100, N_1 = 1000$

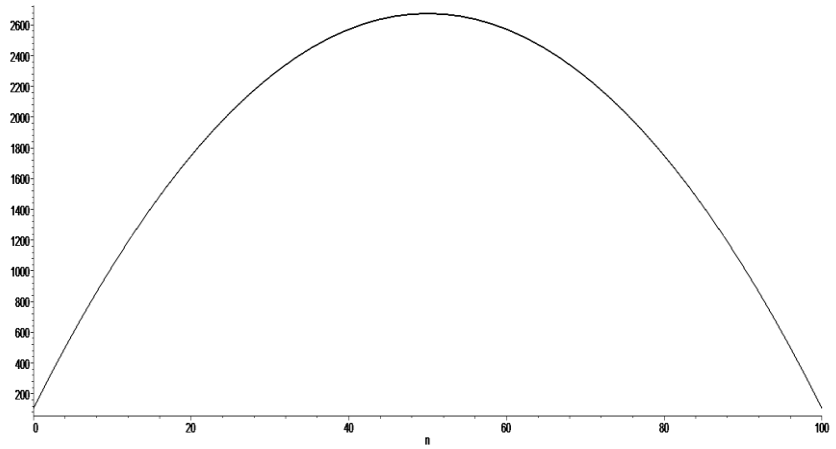

Fig. 1 : graphe de la variance en fonction de n

Si n, N_1 sont fixés, la variance décroît avec N ; elle est proportionnelle à $\frac{1}{N}$; voici un exemple pour $n = 10, N_1 = 1000$:

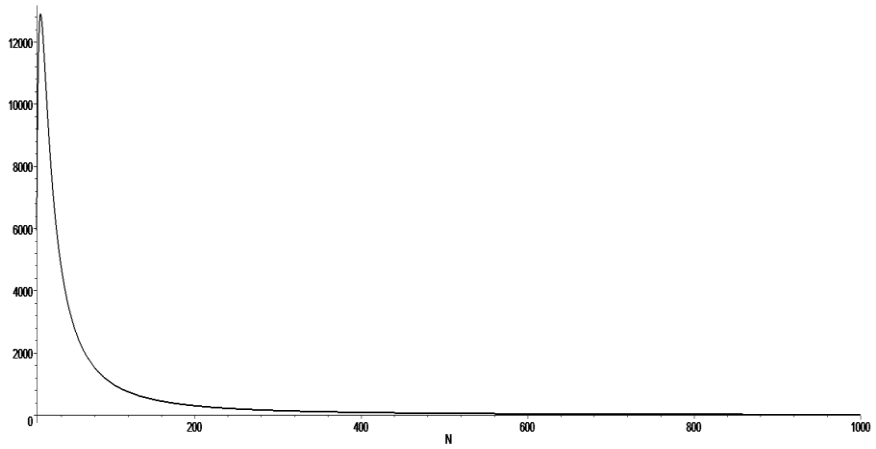

Fig. 2 : graphe de la variance en fonction de N

Enfin, si n, N sont fixés, la variance augmente comme N_1^2 ; voici un exemple pour $n = 10, N = 100$:

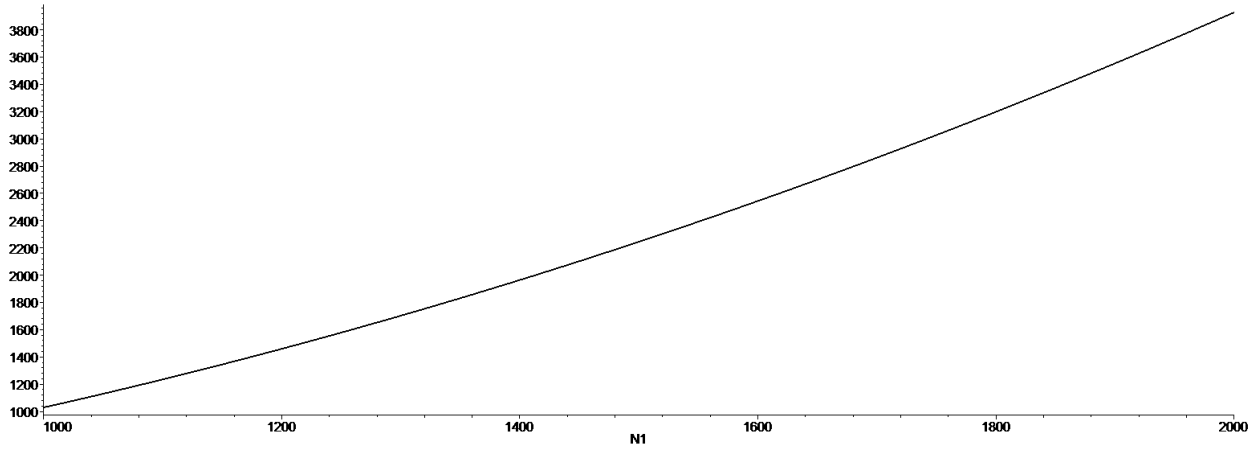

Fig. 3 : graphe de la variance en fonction de N_1

Lorsque le quotient $\alpha = \dfrac{n}{N}$ est fixé, la variance décroît lorsque N augmente, parce qu'on a davantage d'information. Voici la représentation graphique de $P(n_1)$, $n_1 = 0,...,N_1$, dans les deux cas $n = 10, N = 100$ (en rouge) et $n = 20, N = 200$ (en noir) :

Représentation graphique de $P(n_1)$ *dans deux cas*

L'aire avant $\frac{n}{N}$ et l'aire après ne sont pas égales.

Par exemple, pour $n = 10,\ N = 100,\ N_1 = 1000,$

$$\sum_{n_1=0}^{100} P(n_1) = 0.44, \quad \sum_{n_1=101}^{1000} P(n_1) = 0.56.$$

6. Comparaison avec une loi binomiale

On peut vouloir comparer la loi $P(n_1)$ à la loi binomiale :

$$f(n_1) = \binom{N_1}{n_1}\left(\frac{n}{N}\right)^{n_1}\left(1-\frac{n}{N}\right)^{N_1-j}$$

qui prend son maximum au même point, à savoir $\frac{n}{N}N_1$

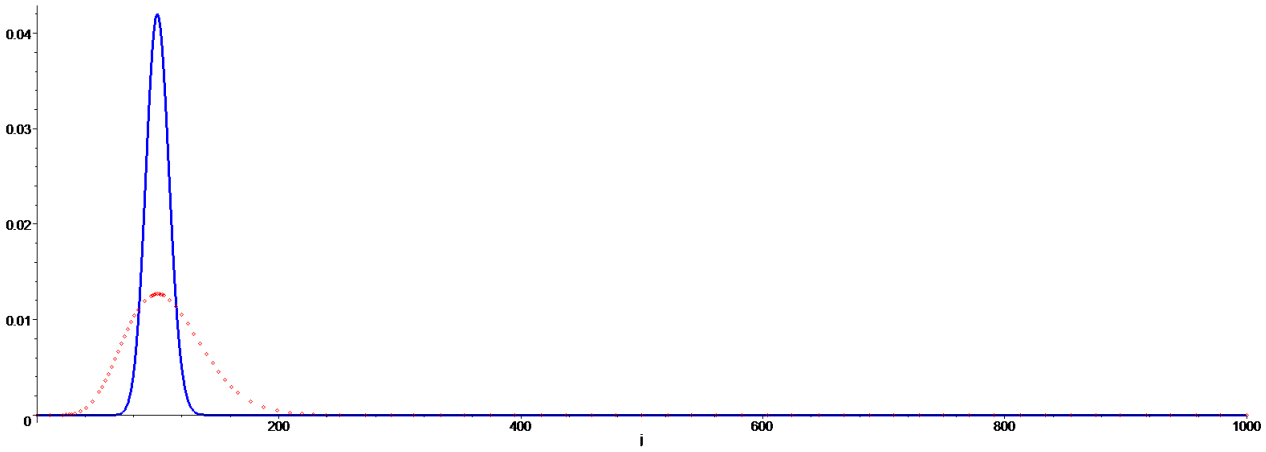

Fig. 4 : comparaison à la loi binomiale

Ici $n=10, N=100, N_1=1000.$ En rouge, $p(n_1)$, en bleu $f(n_1)$; la loi binomiale est beaucoup plus concentrée.

La loi $P(n_1)$ ne peut être approximée par une loi binomiale de paramètres quelconques, $B(M,q)$: l'espérance d'une telle loi est $E_1 = Mq$ et la variance $V_1 = Mq(1-q)$. La loi $P(n_1)$ dépend de trois paramètres (n, N, N_1), tandis que la loi binomiale ne dépend que de deux paramètres. La variance de la loi $P(n_1)$:

$$V = \frac{(n+1)N_1(N+N_1)(N-n)}{N^3}$$

ne se laisse pas mettre sous la forme d'une expression ne dépendant que de $\frac{n}{N}, N_1$.

7. Détermination d'un intervalle de confiance

L'une des questions auxquelles répond l'outil EvalRisk est la détermination d'un intervalle de confiance à 95% ; plus précisément, il calcule n_1' tel que $\sum_{n_1 \le n_1'} P(n_1) \approx 0.025$ et n_1'' tel que $\sum_{n_1 \ge n_1'} P(n_1) \approx 0.025$: encadrement symétrique, laissant une probabilité de 95% au milieu.

On peut se demander s'il est possible d'utiliser l'inégalité de Bienaymé-Chebycheff pour obtenir un tel intervalle de confiance. Cette inégalité se présente sous la forme :

$$P\left(\left|X - E(X)\right| > \alpha\right) < \frac{V}{\alpha^2}$$

et ici, puisque l'espérance et la variance sont connues de manière explicite :

$$P\left(\left|X-\frac{(n+1)N_1}{N+2}\right|>\alpha\right)<\frac{(n+1)N_1(N+N_1+2)(N-n+1)}{(N+2)^2(N+3)\alpha^2}$$

Si on veut que :

$$P\left(\left|X-\frac{(n+1)N_1}{N+2}\right|>\alpha\right)<0.05$$

Il faut choisir α pour que :

$$\frac{(n+1)N_1(N+N_1+2)(N-n+1)}{(N+2)^2(N+3)\alpha^2}=0.05$$

d'où approximativement :

$$\alpha=\sqrt{\frac{n(N-n)N_1(N+N_1)}{N^3 0.05}}$$

Avec $n=10, N=100, N_1=1000,$ on trouve $\alpha\approx 141$; l'intervalle de confiance a donc pour largeur 242 et l'estimation est mauvaise. Cela tient au fait que la densité de probabilité $P(n_1)$ n'est pas symétrique par rapport à sa valeur maximale :

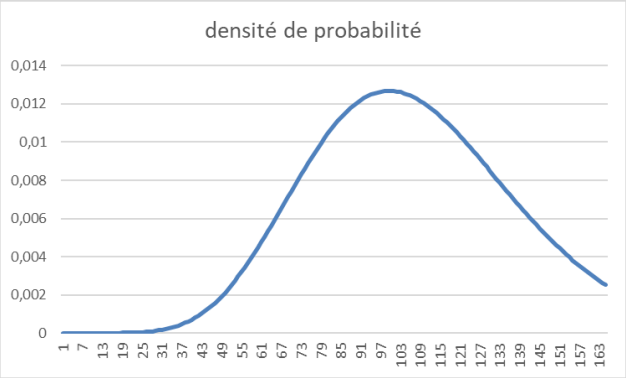

Fig. 5 :graphe de $P(n_1)$ au voisinage du maximum

Toujours dans le cas $n=10, N=100, N_1=1000,$ l'outil EvalRisk donne les bornes $n_1'=53,$ $n_1''=178$; la largeur de l'intervalle est 155 et l'estimation est bien meilleure.

L'outil EvalRisk ne fait pas appel à l'inégalité de Bienaymé-Chebycheff. Il parcourt la loi $P(n_1)$ en commençant à $n_1=0,$ cumule les $P(n_1),$ détermine n_1' lorsque le cumul atteint 0.025 puis n_1'' lorsque le cumul atteint 0.975.

Le programme de calcul, en VBA

```
Option Explicit
Sub macro1()

Sheets(1).Cells(14, 4) = "Calcul en cours"
Sheets(1).Cells(18, 2) = ""
Dim cp As Double 'cumul probas
Dim epsilon1 As Double
epsilon1 = 0.025
Dim epsilon2 As Double
epsilon2 = 0.975
Dim p As Double
Dim n1 As Long
Dim n As Long
n = Sheets(1).Cells(8, 2)
Dim Ntot As Long
Ntot = Sheets(1).Cells(7, 2)
Dim Ntot1 As Long
Ntot1 = Sheets(1).Cells(9, 2)
Dim dL As Long 'défauts lot
dL = Sheets(1).Cells(10, 2)
Dim lp As Double
'initialisation de lp
Dim k As Long
For k = 1 To n + 1
lp = lp + Log(Ntot - n + k) - Log(Ntot + Ntot1 - n + k)
Next k

While cp < epsilon1
lp = lp + Log(n + n1 + 1) + Log(Ntot1 - n1) - Log(n1 + 1) - Log(Ntot + Ntot1 - n - n1)
p = Exp(lp)
cp = cp + p
n1 = n1 + 1
If n1 = dL Then Sheets(1).Cells(16, 2) = 1 - cp
Wend
Sheets(1).Cells(14, 2) = n1

While cp < epsilon2
lp = lp + Log(n + n1 + 1) + Log(Ntot1 - n1) - Log(n1 + 1) - Log(Ntot + Ntot1 - n - n1)
p = Exp(lp)
cp = cp + p
n1 = n1 + 1
If n1 = dL Then Sheets(1).Cells(10, 2) = 1 - cp
Wend
Sheets(1).Cells(15, 2) = n1
End Sub
```