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

Commentaires sur le livre de Laplace

"Théorie Analytique des Probabilités", 1820

par Bernard Beauzamy

avril 2015

I. Evaluation d'un taux de risque

Laplace mentionne explicitement la formule :

$$p = \frac{N+1}{N+2}$$

comme étant la probabilité qu'un événement s'étant produit N fois sur N expériences se produise encore à la $N+1$ème expérience. Mais l'argumentation qu'il utilise est très difficile à comprendre ; voici le raisonnement fait (Introduction, page XVII ; sauf erreur de notre part, le raisonnement n'est pas détaillé ailleurs) :

"Quand la probabilité d'un événement simple est inconnue, on peut lui supposer également toutes les valeurs depuis zéro jusqu'à l'unité. La probabilité de chacune de ces hypothèses, tirée de l'événement observé, est, par le sixième principe, une fraction dont le numérateur est la probabilité de l'événement dans cette hypothèse, et dont le dénominateur est la somme des probabilités semblables relatives à toutes les hypothèses. Ainsi la probabilité que la possibilité de l'événement est comprise dans des limites données est la somme des fractions comprises dans ces limites. Maintenant, si l'on multiplie chaque fraction par la probabilité de l'événement futur, déterminée dans l'hypothèse correspondante, la somme des produits relatifs à toutes ces hypothèses sera, par le septième principe, la probabilité de l'événement futur, tirée de l'événement observé. On trouve ainsi qu'un événement étant arrivé de suite un nombre quelconque de fois, la probabilité qu'il arrivera encore la fois suivante est égale à ce nombre augmenté de l'unité, divisé par le même nombre augmenté de deux unités."

Essayons de clarifier ceci.

Soit X une variable aléatoire binaire (valeurs 0 ou 1) ; la loi de X est inconnue. Notons $p = P(X = 0)$, qui est inconnu. On sait que sur N répétitions la valeur 0 est sortie à chaque fois ; il s'agit d'estimer p. Notons TR (Taux de Risque) la valeur de p, considérée comme variable aléatoire, dont il s'agit précisément d'estimer la loi. A priori, comme dit Laplace, puisque la loi de TR est inconnue, "on peut lui supposer également toutes les valeurs depuis 0 jusqu'à l'unité", ce qui revient à dire que l'on peut lui attribuer une loi uniforme sur l'intervalle $[0,1]$. Pour simplifier le raisonnement, supposons d'abord que TR ne prenne que les valeurs discrètes $k/10$, $k = 0,...,10$.

Supposons que sur N essais on ait n réalisations de l'événement, ce que nous noterons en abrégé "n sur N" (dans le cas présent, $n = N$). Alors, on peut écrire formellement :

$$P\left(TR = \frac{k}{10} \mid n \text{ sur } N\right) = \frac{P\left(TR = \frac{k}{10} \text{ et } n \text{ sur } N\right)}{P(n \text{ sur } N)}$$
$$= \frac{P\left(n \text{ sur } N \mid TR = \frac{k}{10}\right) P\left(TR = \frac{k}{10}\right)}{P(n \text{ sur } N)}$$

Or on a aussi :

$$P(n \text{ sur } N) = \sum_{j=0}^{10} P\left(n \text{ sur } N \text{ et } TR = \frac{j}{10}\right)$$
$$= \sum_{j=0}^{10} P\left(n \text{ sur } N \mid TR = \frac{j}{10}\right) P\left(TR = \frac{j}{10}\right)$$

et par conséquent :

$$P\left(TR = \frac{k}{10} \mid n \text{ sur } N\right) = \frac{P\left(n \text{ sur } N \mid TR = \frac{k}{10}\right) P\left(TR = \frac{k}{10}\right)}{\sum_{j=0}^{10} P\left(n \text{ sur } N \mid TR = \frac{j}{10}\right) P\left(TR = \frac{j}{10}\right)} \quad (1)$$

Or, dans cette présentation, la loi a priori de TR est une loi uniforme ; autrement dit, $P\left(TR = \frac{k}{10}\right)$ a la même valeur pour tout k. La formule ci-dessus se simplifie donc :

$$P\left(TR = \frac{k}{10} \mid n \text{ sur } N\right) = \frac{P\left(n \text{ sur } N \mid TR = \frac{k}{10}\right)}{\sum_{j=0}^{10} P\left(n \text{ sur } N \mid TR = \frac{j}{10}\right)} \quad (2)$$

On obtient ce que dit Laplace : "une fraction dont le numérateur est la probabilité de l'événement dans cette hypothèse, et dont le dénominateur est la somme des probabilités semblables relatives à toutes les hypothèses".

Nous cherchons la probabilité d'avoir un succès au $N+1$ème essai, sachant que nous avons eu n succès sur N essais. Cette quantité est donnée par la formule :

$$q = \sum_{k=0}^{10} \frac{k}{10} P\left(TR = \frac{k}{10} \,|\, n \text{ sur } N \right) \tag{3}$$

soit encore :

$$q = \frac{\sum_{k=0}^{10} \frac{k}{10} P\left(n \text{ sur } N \,|\, TR = \frac{k}{10} \right)}{\sum_{j=0}^{10} P\left(n \text{ sur } N \,|\, TR = \frac{j}{10} \right)} \tag{4}$$

Revenons à des notations continues, plus simples à manipuler à ce stade. On a, d'après la formule précédente :

$$q = \frac{\int_0^1 \lambda P\left(n \text{ sur } N \,|\, TR = \lambda\right) d\lambda}{\int_0^1 P\left(n \text{ sur } N \,|\, TR = \lambda\right) d\lambda} \tag{5}$$

Si $TR = \lambda$, la probabilité d'avoir n succès sur N essais est donnée par la formule du binôme :

$$P\left(n \text{ sur } N \,|\, TR = \lambda\right) = \binom{N}{n} \lambda^n \left(1-\lambda\right)^{N-n} \tag{6}$$

et donc :

$$q = \frac{\int_0^1 \lambda^{n+1} \left(1-\lambda\right)^{N-n} d\lambda}{\int_0^1 \lambda^n \left(1-\lambda\right)^{N-n} d\lambda} \tag{7}$$

Si $n = N$, on obtient :

$$q = \frac{\int_0^1 \lambda^{N+1} d\lambda}{\int_0^1 \lambda^N d\lambda} = \frac{N+1}{N+2}, \tag{8}$$

formule effectivement annoncée par Laplace.

La formule (2) est un peu difficile à comprendre ; la simplification à partir de (1) résulte du fait que, à ce stade, on admet la loi uniforme sur TR, alors que la suite des calculs montrera que TR est très proche de 1, lorsque l'on a N succès sur N essais. La formule (2) ne résulte pas d'une formule générale du type :

$$P(T=k) = \frac{P(X=0 \mid T=k)}{\sum_j P(X=0 \mid T=j)} \tag{*}$$

pour deux variables aléatoires quelconques X, T. La formule (*) est fausse, comme le montre l'exemple très simple de deux variables ayant la loi conjointe suivante :

X\T	0	1
0	1/2	1/4
1	1/6	1/12

Rappelons que dans le livre [NMP], on peut trouver la démonstration du fait que, si l'on a n succès sur N essais, le taux de risque a pour densité :

$$f_{n,N}(\lambda) = \frac{(N+1)!}{n!N!} \lambda^n (1-\lambda)^{N-n} \tag{9}$$

d'où la formule (8) résulte immédiatement.

II. Inégalités du nombre garçons-filles à la naissance

Dans son livre, Laplace dit que l'on observe en moyenne 22 naissances de garçons pour 21 filles et il montre que cette divergence par rapport à l'équiprobabilité ne peut pas être due au hasard.

Les chiffres modernes sont étonnamment proches :

France métropolitaine				
source insee				
année	Total	Garçons	Filles	g/(g+f)
2003	761 464	389 349	372 115	0,51131636
2004	767 816	393 477	374 339	0,51246262
2005	774 355	396 346	378 009	0,51184018
2006	796 896	407 846	389 050	0,51179326
2007	785 985	402 297	383 688	0,51183801
2008	796 044	406 784	389 260	0,51100693
2009	793 420	405 902	387 518	0,51158529
2010	802 224	410 140	392 084	0,51125371
2011	792 996	405 206	387 790	0,51098114
2012	790 290	404 774	385 516	0,51218413
2013	781 621	400 149	381 472	0,51194761

Total sur 11 années :
nombre total de naissances 8 643 111
nombre de garçons : 4 422 270
ratio garçons/total : 0,511652575

Pour Laplace, ratio 22/43 = 0,5116279.

Le ratio calculé par Laplace (sur un petit nombre d'années) est intermédiaire entre ceux constatés sur la période 2003-2013.

Démontrons, comme le fait Laplace, que la différence par rapport à 1/2 ne peut être due au hasard.

Notons X_k la variable aléatoire qui vaut 0 si le $k-ème$ enfant est un garçon, 1 s'il est une fille. Nous avons $N = 8\,643\,111$ observations.

Faisons les hypothèses suivantes :

a) les X_k sont indépendantes ;

b) les X_k ont toutes la même loi ;

c) $P(X_k = 0) = P(X_k = 1) = \frac{1}{2}$

Alors les X_k ont pour moyenne $m = \frac{1}{2}$ et pour écart-type $\sigma = \frac{1}{2}$. On peut écrire, pour tout $\delta > 0$:

$$P(Z > \delta) = \int_{\delta}^{+\infty} \exp\left(-t^2/2\right) \frac{dt}{\sqrt{2\pi}} \qquad (1)$$

où $Z = \frac{\frac{1}{N}\sum_1^N X_k - m}{\sigma / \sqrt{N}}$; en effet, cette variable suit une loi normale avec une très bonne approximation, sous les hypothèses ci-dessus.

La condition (1) s'écrit encore :

$$P\left(\frac{1}{N}\sum_1^N X_k > m + \frac{\delta\sigma}{\sqrt{N}}\right) = \int_{\delta}^{+\infty} \exp\left(-t^2/2\right) \frac{dt}{\sqrt{2\pi}} \qquad (2)$$

et, avec les valeurs numériques ci-dessus et $\delta = 50,\ m + \frac{\delta\sigma}{\sqrt{N}} = 0.508$, la probabilité ci-dessus vaut $\int_{50}^{+\infty} \exp\left(-t^2/2\right) \frac{dt}{\sqrt{2\pi}}$, qui est infime. Il est donc "virtuellement impossible", sous les hypothèses ci-dessus, que la moyenne des variables prenne la valeur 0.511. Comme elle le fait (et sur de très longues périodes), on en déduit que l'une au moins des hypothèses ci-dessus est fausse. Le problème est que l'on ne sait pas laquelle, et Laplace ne discute pas cette question.

On peut raisonnablement penser que la loi est stationnaire, c'est-à-dire que les probabilités sont les mêmes dans le temps (voir graphique ci-dessous). Il reste donc deux possibilités :

- Ou bien les variables ne sont pas indépendantes : il se pourrait par exemple qu'une famille avec un enfant mâle ait un tout petit peu plus de chances, pour le second enfant, d'avoir un garçon ;

- Ou bien la loi n'est pas équiprobable : il y a un tout petit peu plus de chances, pour chaque naissance, d'avoir un garçon plutôt qu'une fille.

La formule (9) du paragraphe précédent nous donne l'expression de la densité du taux de risque (ici le "risque" est d'avoir un garçon), avec $n = 4\ 422\ 270$ et $N = 8\,643\,111$. L'espérance vaut $q = \frac{n+1}{N+2} \simeq 0.5116525724.$

Voici l'évolution du quotient $\frac{g}{g+f}$ sur une longue période :

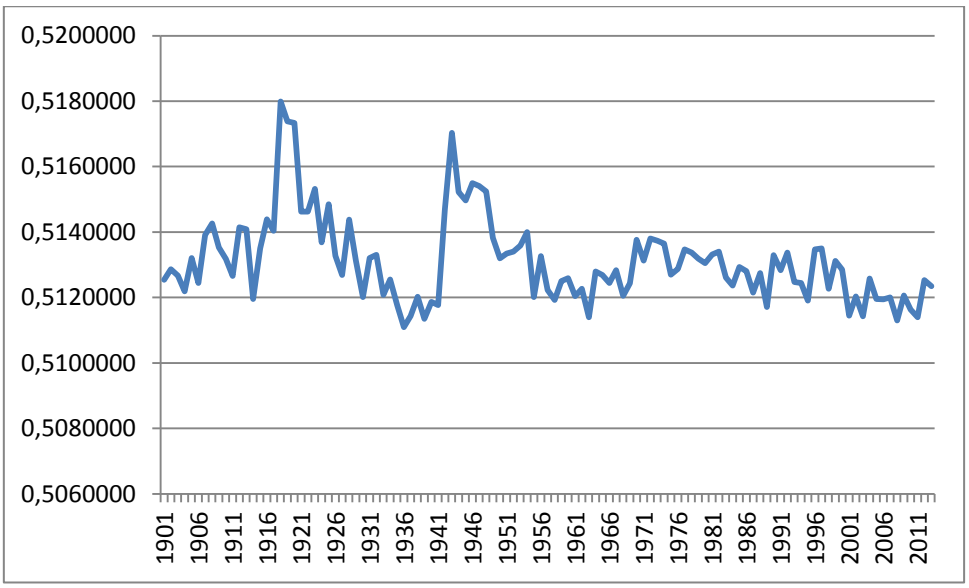

Figure 1 : évolution du ratio g/(g+f) depuis 1900

Source : INSEE, statistiques de l'état civil

Le graphique ci-dessus concerne tous les enfants nés (vivants ou sans vie), sauf pour les trois années 2000, 2001, 2002, où le nombre d'enfants nés sans vie n'était pas disponible (il s'agit alors simplement du nombre d'enfants vivants).

L'INSEE fait les commentaires suivants :

- Les fortes évolutions du nombre d'enfants sans vie en 2002, puis en 2008 et 2009, sont liées à des changements législatifs (voir rubrique Documentation, "les sources des statistiques de l'état civil").
- En particulier, depuis août 2008, les anciens critères de durée d'aménorrhée et de poids ne sont plus pris en compte et les données françaises récentes sur les enfants sans vie ne peuvent pas être comparées à celles des autres pays.
- Depuis cette date, les déclarations d'enfants sans vie à l'état civil reposent sur une démarche volontaire des parents.
- La répartition des enfants sans vie par sexe n'est pas disponible de 2000 à 2002.

On constate une forte croissance de la proportion de garçons après les guerres.

Un "principe de Fisher" (voir par exemple Wikipedia http://fr.wikipedia.org/wiki/Sex-ratio) postule que, pour la plupart des espèces, le sex-ratio (défini comme le quotient g/f) est approximativement de 1.1 et donne à ce déséquilibre une explication de nature "stratégique", qui ne nous paraît pas convaincante.

En effet, les chiffres ci-dessus, qui reflètent un déséquilibre g/f, ne concernent que les naissances. Entre la fécondation et la naissance, il y a la période de grossesse, qui peut comporter une fausse couche ou un avortement. On estime à 20% des grossesses celles qui donnent lieu à une fausse couche (nous n'avons pas les chiffres exacts) ; pour environ 750 000 naissances par an, il faudrait donc environ 937 500 fécondations, résultant en 187 500 fausses couches.

En ce qui concerne l'IVG (interruption volontaire de grossesse), leur nombre est de 222 500 en France pour l'année 2010. Les chiffres sont évidemment inconnus avant la légalisation de l'avortement. Au total, on a un nombre approximatif de 400 000 enfants conçus mais non nés, dont le sexe est inconnu. Ce nombre est considérable, et peut entièrement fausser les statistiques sur le sex ratio à la naissance.

Il suffirait que les fausses couches et/ou les avortements touchent légèrement plus les filles que les garçons pour expliquer le déséquilibre du sex ratio que nous constatons. A la suite des guerres, comme les conditions de vie sont très précaires, le nombre de fausses couches peut parfaitement augmenter, corroborant ainsi l'explication précédente.

Nos conclusions sont donc les suivantes :

- Il y a clairement un déséquilibre statistiquement significatif à la naissance : plus de garçons que de filles ;

- Néanmoins, le pourcentage élevé d'enfants qui sont conçus mais meurent avant la naissance (environ 400 000 sur 750 000) ne permet pas de conclure à un déséquilibre H/F lors de la conception, puisque le sexe de ces enfants est ignoré par les statistiques ;

- Les théories qui prétendent que le déséquilibre H/F à la naissance est voulu par la nature et résulterait d'une adaptation stratégique nous paraissent entièrement dépourvues de valeur scientifique, dans la mesure où les enfants conçus mais non nés ne sont pas pris en considération.

III. Référence

[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.