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

Commentaires sur le livre de Laplace

"Théorie Analytique des Probabilités", 1820

et analyse du "sex-ratio" selon Laplace

par Bernard Beauzamy

I. Eléments historiques

Laplace (1749-1827) est habituellement considéré comme le fondateur de la Théorie des Probabilités, même si les calculs de probabilités sont évidemment beaucoup plus anciens : les Phéniciens assuraient la cargaison de leurs navires et les calculs sur les jeux de hasard ont existé de tous temps.

Les probabilités se caractérisent par une extrême simplicité dans l'énoncé des problèmes et une très grande difficulté dans leur résolution. Un exemple frappant est donné par le jeu de pile ou face : si, à chaque partie, le perdant donne un euro au gagnant, la fortune de chaque joueur, au bout de la n-ème partie, est presque sûrement bornée par la "courbe de Khintchine" :

$$y = \pm\sqrt{2nLog\left(Log\left(n\right)\right)}$$

et, inversement, il est presque sûr que la fortune de chaque joueur atteindra cette courbe, à un moment ou à un autre. Le sens des mots "presque sûr" est déjà peu clair en soi. Pour une version quantitative un peu clarifiée et simplifiée, voir notre livre "Simple Random Walks" [SRW].

Laplace est connu pour son indépendance d'esprit ; lorsqu'il présente son "Traité de Mécanique céleste" à Napoléon 1er, celui-ci lui demande : "où est Dieu dans tout cela ?" et Laplace répond : "Sire, je n'avais pas besoin de cette hypothèse". De nos jours, la réponse de l'immense majorité des scientifiques serait : "Sire, puisque vous le souhaitez, je vais vous démontrer que la planète est en danger".

Nous nous intéressons ici au calcul d'un "taux de risque" : sachant qu'un événement s'est produit n fois sur N expériences passées, quelle est la probabilité qu'il se produise n' fois sur N' expériences futures ? Là encore, énoncé simple, réponse difficile.

II. 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 :

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

et par conséquent :

$$P\left(TR=\frac{k}{10}\,|\,n\ sur\ N\right)=\frac{P\left(n\ sur\ N\,|\,TR=\frac{k}{10}\right)P\left(TR=\frac{k}{10}\right)}{\sum_{j=0}^{10}P\left(n\ sur\ N\,|\,TR=\frac{j}{10}\right)P\left(TR=\frac{j}{10}\right)}\qquad(1)$$

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 :

$$P\left(TR=\frac{k}{10}\,|\,n\ sur\ N\right)=\frac{P\left(n\ sur\ N\,|\,TR=\frac{k}{10}\right)}{\sum_{j=0}^{10}P\left(n\ sur\ N\,|\,TR=\frac{j}{10}\right)}\qquad(2)$$

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 :

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

soit encore :

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

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

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

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

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

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 :

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

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

Dans notre 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 :

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

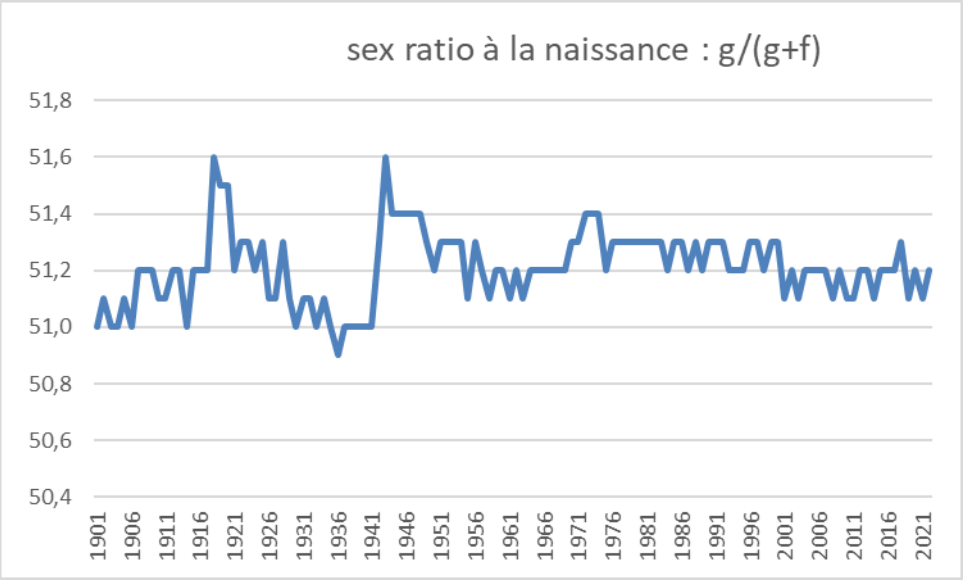
d'où la formule (8) résulte immédiatement.

III. 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.

1. Données numériques

Les chiffres modernes sont étonnamment proches :



Source : Insee, statistiques de l'état civil

	Ensemble des nés vivants	Nés vivants - Garçons	Nés vivants - Filles	Garçons vivants pour 100 nés vivants
1901	917 075	468 125	448 950	51,0
1902	904 434	462 097	442 337	51,1
1903	884 498	451 510	432 988	51,0
1904	877 091	447 651	429 440	51,0
1905	865 604	442 397	423 207	51,1
1906	864 745	441 358	423 387	51,0
1907	829 632	424 692	404 940	51,2
1908	848 982	435 027	413 955	51,2
1909	824 739	421 882	402 857	51,2
1910	828 140	423 581	404 559	51,1
1911	793 506	405 239	388 267	51,1
1912	801 642	410 816	390 826	51,2
1913	795 851	407 569	388 282	51,2
1914	757 931	386 560	371 371	51,0
1915	482 968	247 205	235 763	51,2
1916	384 676	197 133	187 543	51,2
1917	412 744	211 353	201 391	51,2
1918	472 816	244 107	228 709	51,6
1919	506 960	261 049	245 911	51,5
1920	838 137	432 044	406 093	51,5
1921	816 555	418 461	398 094	51,2
1922	764 373	391 800	372 573	51,3
1923	765 888	393 194	372 694	51,3
1924	757 873	387 889	369 984	51,2
1925	774 455	397 044	377 411	51,3
1926	771 690	394 627	377 063	51,1
1927	748 102	382 256	365 846	51,1
1928	753 570	386 216	367 354	51,3
1929	734 140	375 448	358 692	51,1
1930	754 020	384 658	369 362	51,0
1931	737 611	376 921	360 690	51,1
1932	726 299	371 330	354 969	51,1
1933	682 394	347 973	334 421	51,0
1934	681 518	347 967	333 551	51,1
1935	643 870	328 346	315 524	51,0
1936	634 344	322 982	311 362	50,9
1937	621 453	316 925	304 528	51,0
1938	615 582	314 214	301 368	51,0
1939	615 599	313 712	301 887	51,0
1940	561 281	286 361	274 920	51,0
1941	522 261	266 348	255 913	51,0
1942	575 261	295 346	279 915	51,3
1943	615 780	317 654	298 126	51,6
1944	629 878	323 710	306 168	51,4
1945	645 899	331 721	314 178	51,4
1946	843 904	433 825	410 079	51,4
1947	870 472	447 510	422 962	51,4
1948	870 836	447 898	422 938	51,4
1949	872 661	447 510	425 151	51,3
1950	862 310	441 795	420 515	51,2
1951	826 722	423 797	402 925	51,3
1952	822 204	421 546	400 658	51,3
1953	804 696	412 599	392 097	51,3
1954	810 754	416 078	394 676	51,3
1955	805 917	412 159	393 758	51,1
1956	806 916	413 546	393 370	51,3
1957	816 467	417 634	398 833	51,2
1958	812 215	415 221	396 994	51,1
1959	829 249	424 405	404 844	51,2
1960	819 819	419 775	400 044	51,2
1961	838 633	428 877	409 756	51,1
1962	832 353	425 919	406 434	51,2
1963	868 876	443 844	425 032	51,1
1964	877 804	449 511	428 293	51,2
1965	865 688	443 390	422 298	51,2
1966	863 527	442 128	421 399	51,2
1967	840 568	430 641	409 927	51,2
1968	835 796	427 623	408 173	51,2
1969	842 245	431 346	410 899	51,2
1970	850 381	436 599	413 782	51,3
1971	881 284	451 978	429 306	51,3
1972	877 506	450 667	426 839	51,4
1973	857 186	440 190	416 996	51,4
1974	801 218	411 439	389 779	51,4
1975	745 065	381 804	363 261	51,2
1976	720 395	369 439	350 956	51,3
1977	744 744	382 337	362 407	51,3
1978	737 062	378 281	358 781	51,3
1979	757 354	388 604	368 750	51,3
1980	800 376	410 547	389 829	51,3
1981	805 483	413 480	392 003	51,3
1982	797 223	409 205	388 018	51,3
1983	748 525	383 659	364 866	51,3
1984	759 939	389 310	370 629	51,2
1985	768 431	394 112	374 319	51,3
1986	778 468	399 199	379 269	51,3
1987	767 828	393 231	374 597	51,2
1988	771 268	395 439	375 829	51,3
1989	765 473	391 649	373 824	51,2
1990	762 407	391 312	371 095	51,3
1991	759 056	389 239	369 817	51,3
1992	743 658	381 744	361 914	51,3
1993	711 610	364 589	347 021	51,2
1994	710 993	364 277	346 716	51,2
1995	729 609	373 409	356 200	51,2
1996	734 338	377 003	357 335	51,3
1997	726 768	373 157	353 611	51,3
1998	738 080	378 075	360 005	51,2
1999	744 791	382 132	362 659	51,3
2000	774 782	397 352	377 430	51,3
2001	770 945	394 297	376 648	51,1
2002	761 630	389 981	371 649	51,2
2003	761 464	389 349	372 115	51,1
2004	767 816	393 477	374 339	51,2
2005	774 355	396 346	378 009	51,2
2006	796 896	407 846	389 050	51,2
2007	785 985	402 297	383 688	51,2
2008	796 044	406 784	389 260	51,1
2009	793 420	405 902	387 518	51,2
2010	802 224	410 140	392 084	51,1
2011	792 996	405 206	387 790	51,1
2012	790 290	404 774	385 516	51,2
2013	781 621	400 149	381 472	51,2
2014	781 167	399 284	381 883	51,1
2015	760 421	389 181	371 240	51,2
2016	744 697	381 310	363 387	51,2
2017	730 242	373 716	356 526	51,2
2018	719 737	369 121	350 616	51,3
2019	714 029	364 924	349 105	51,1
2020	696 664	356 389	340 275	51,2
2021	701 819	358 833	342 986	51,1
2022	686 564	351 296	335 268	51,2

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

Le ratio calculé par Laplace (sur un petit nombre d'années) est intermédiaire entre ceux constatés sur la période 1901-2022.

2. Analyse du problème

Démontrons, comme le fait Laplace, que la différence par rapport à 1/2 ne peut être due au hasard. Mais, aujourd'hui, nous serons plus circonspects dans le vocabulaire : nous conclurons que : a) ou bien le sexe n'est pas indépendant d'une naissance à l'autre, b) ou bien il n'y a pas équiprobabilité à chaque naissance.

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 $N = 92\ 764\ 826$ naissances, donc autant d'observations, et $N_F = 45\ 257\ 691$ filles.

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 δ :

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

où $Z = \dfrac{\frac{1}{N}\sum_1^N X_k - m}{\sigma / \sqrt{N}}$; en effet, cette variable suit une loi normale avec une très bonne approximation, sous les hypothèses ci-dessus. Avec les valeurs numériques ci-dessus, $Z \approx -233.6$.

Or la probabilité qu'une variable suivant la loi normale prenne une valeur inférieure à -20 est inférieure à 10^{-88}. 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-dessus). 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 $N_G = N - N_F = 47\ 507\ 135$ et $N = 92\ 764\ 826$. L'espérance vaut $q = \frac{N_G + 1}{N + 2} \simeq 0.512$.

3. Remarques complémentaires

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

1918	51,6
1919	51,5
1920	51,5
1943	51,6
1944	51,4
1945	51,4
1946	51,4
1947	51,4
1948	51,4

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 216 000 en France pour l'année 2022. 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.

IV. Critiques complémentaires

Laplace dit : "Une intelligence qui, à un instant donné, connaîtrait toutes les forces dont la nature est animée, la position respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'Univers, et ceux du plus léger atome. Rien ne serait incertain pour elle, et l'avenir comme le passé seraient présents à ses yeux" (Essai philosophique sur les probabilités, Paris, Bachelier, 1840).

La physique moderne, depuis la mécanique quantique, est certainement en désaccord avec une telle affirmation. On peut penser que Laplace l'a formulée alors qu'il travaillait sur les orbites des planètes, mais, même dans ce cas, elle est sujette à caution. On ne sait toujours pas démontrer que le système solaire est stable, ni même l'orbite de la Lune. Il est théoriquement possible que des perturbations infimes (le choc d'un météorite, par exemple) arrachent une planète à son orbite, l'envoyant soit vers le soleil, soit vers l'espace extérieur.

Mais, par-dessus tout, Laplace était bien placé pour savoir que même la vaste intelligence dont il parle serait bien incapable de prévoir le résultat d'une partie de pile ou face. La conclusion purement déterministe du principal fondateur du calcul des probabilités est donc étonnante en soi.

V. Références

[SRW] Bernard Beauzamy : Simple Random Walks in the Plane, in English. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN : 979-10-95773-01-6, ISSN: 1767-1175. Relié, 208 pages. Février 2020.

[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.