1
ÉTUDES

Garçon-fille au hasard ?

par Jean-Pierre ROSAY, Université de Provence

Rédigé plutôt sous forme de problème, le texte suivant est une adaptation d'un sujet d'examen de DEUG B (Marseille, juin 1980), lui-même basé sur des chiffres fournis dans le livre "Bases Statistiques" de F. GREMY et D. SALMON (Sciences Mathématiques au service de la médecine, Dunod, 1969), p. 67.

Il s'agit d'examiner l'hypothèse selon laquelle, dans une même famille, il y a, à chaque naissance, une même probabilité P d'avoir un garçon (donc qu'en particulier le sexe des enfants déjà nés est sans influence sur le sexe d'un enfant à naître) ; cette hypothèse sera énoncée plus précisément dans le paragraphe 2.

Il se trouve qu'une telle hypothèse peut être rejetée (au seuil 5 %) avec un outillage mathématique élémentaire, *entièrement à la portée des élèves des classes terminales,* en particulier sans recours au théorème de Gauss ou central limite. On a ainsi, en restant au niveau de classe terminale, la possibilité de donner une toute première initiation aux tests statistiques.

0. Paragraphe préliminaire. Calcul des moments des lois binomiales

Soient $n \in \mathbf{N}_*$, $p \in [0;1]$ et X une variable aléatoire de loi binomiale B_n^p . C'est-à-dire, si $k \in \{0,1,\ldots,n\}$,

$$\text{Prob}\,\{X=k\} = C_n^k\, p^k\, (1-p)^{n-k}$$

Soit $r \in \mathbf{N}$; le moment d'ordre r de X par rapport à sa moyenne est l'espérance mathématique (ou moyenne) de $(X-np)^r$, c'est-à-dire le nombre

$$E[(X-np)^r] = \sum_{k=0}^{n} (k-np)^r\, C_n^k\, p^k\, (1-p)^{n-k}$$

Pour $p \in [0;1]$, définissons

$$F_r(p) = \sum_{k=0}^{n} (k-np)^r C_n^k p^k (1-p)^{n-k} ;$$

par dérivation, on obtient

$$F_r'(p) = -n r F_{r-1}(p) + \frac{1}{p(1-p)} F_{r+1}(p)$$

d'où la formule de récurrence :

$$F_{r+1}(p) = p(1-p)[F_r'(p) + n r F_{r-1}(p)]$$

Partant du fait évident $F_0(p) = 1$, on tire :

$E(X-np) = F_1(p) = 0$ (on retrouve ainsi le fait que $E(X) = np$)

$E[(X-np)^2] = F_2(p) = n p(1-p)$ (la variance)

$E[(X-np)^3] = F_3(p) = n p(1-p)(1-2p)$

$E[(X-np)^4] = F_4(p) = n^2 3p^2 (1-p)^2 + n p(1-p)(1-6p+6p^2)$

D'où, sans même calculer F_5, on voit que

$$E[(X-np)^6] = F_6(p) = n^3 (15p^3(1-p)^3) + n^2 \ldots + n \ldots$$

etc.

1. Estimation de la probabilité que X, de loi binomiale B_n^p, s'écarte de sa moyenne d'au moins k fois son écart-type

Les notations du paragraphe 0 sont conservées.

Soit $k>0$. Posons $\pi = \text{Prob}\{|X-np| \geqslant k\sigma\}$ où σ est l'écart-type de X, $\sigma = \sqrt{npq}$, et $q = 1-p$.

Il y a deux résultats que "tout le monde connaît", 1.1 et 1.2 :

1.1 Si n est assez grand, π est calculé, approximativement, à l'aide de la loi normale réduite $n(0,1)$: $\pi \simeq (n(0,1),]-\infty, -k] \cup [k, +\infty[)$

Pour $k=2$, on trouve $\pi \simeq 0{,}0456$;

$k=3$, on trouve $\pi \simeq 0{,}0027$.

Ce résultat a évidemment pour inconvénient de ne pas être élémentaire.

1.2 L'inégalité de Bienaymé-Tchebichev donne la majoration $\pi \leqslant \frac{1}{k^2}$.

On voit que cette majoration est fort médiocre ; l'inégalité de Bienaymé-Tchebichev est un outil très général et de ce fait assez grossier.

1.3 Une inégalité meilleure pour certaines valeurs de k peut être déduite du calcul du moment d'ordre 4 $E[(X-np)^4]$, en copiant exactement la démonstration de l'inégalité de Bienaymé-Tchebichev.
On a le lemme suivant :

Lemme

Soit Y *une variable aléatoire* $\geqslant 0$. *Pour tout* $a>0$,

$$E(Y) \geqslant a \text{ Prob } \{Y \geqslant a\} .$$

Evidemment ! Si les deux tiers des personnes d'une assemblée pèsent chacune plus de 100 kilos, le poids moyen des personnes de cette assemblée dépasse $100 \times \frac{2}{3}$.

L'inégalité de Bienaymé-Tchebichev pour X en découle en prenant

$$Y = (X-np)^2 \qquad \text{et} \qquad a = k^2\sigma^2 .$$

Prenons maintenant

$$Y = (X-np)^4 \qquad \text{et} \qquad a = h^4\theta^4$$

$$\text{où} \qquad \theta^4 = E(Y) = 3n^2p^2q^2 + n(pq-6p^2q^2) ;$$

il vient

$$\text{Prob}\{|X-np| > h\theta\} \leqslant \frac{1}{h^4} \qquad \text{pour tout } h>0 .$$

Soit en posant $h\theta = k\sigma$ et donc

$$k = \frac{[3n^2p^2q^2 + n(pq-6p^2q^2)]^{1/4}}{\sqrt{npq}}\, h$$

$$\text{Prob}\{|X-np| > k\sigma\} \leqslant \frac{3 - \frac{6}{n} + \frac{1}{npq}}{k^4}$$

Pour n assez grand,

$$\text{Prob}\{|X-np| > k\sigma\} \leqslant \frac{3}{k^4} \text{ (approximativement),}$$

ce qui donne les majorations suivantes de π :

pour $k = 2$: $\pi \leqslant 0{,}1875$

pour $k = 3$: $\pi \leqslant 0{,}03703\overline{703}$

pour $k = 2{,}35$: $\pi \leqslant 0{,}10 = 2\times 0{,}05$

Cette dernière majoration nous suffirait pour terminer le problème pourvu que nous sachions que, approximativement pour n grand,

$$\text{Prob}\{(X-np) < -k\sigma\} \simeq \text{Prob}\{(X-np) > k\sigma\}$$

Je ne vois toutefois pas de moyen élémentaire d'y parvenir sauf dans le cas de $p = \frac{1}{2}$. Alors passons au moment d'ordre 6.

1.4 Pour n grand, on a donc $\frac{E[(X-np)^6]}{n^3p^3q^3} \simeq 15$,

d'où l'on tire par un raisonnement semblable :

$$\text{Prob}\{|X - np| > k\sigma\} \leqslant \frac{15}{k^4}$$

D'où enfin, pourvu que n soit assez grand,

$$\text{Prob}\{|X - np| > 2{,}6\,\sigma\} \leqslant 0{,}05$$

2. Calcul des probabilités. Etude de la proportion de familles ayant 4 garçons et 4 filles, parmi les familles de 8 enfants

Soit X le nombre de familles de 4 garçons et 4 filles parmi n familles de 8 enfants prises au hasard. On suppose n grand.

Faisons une "prédiction", en nous basant successivement sur les trois hypothèses suivantes :

2.1 ***Hypothèse 1*** :

"Lors de toute naissance, il y a probabilité $\frac{1}{2}$ de naissance d'un garçon". La probabilité pour que parmi les 8 enfants d'une famille il y ait exactement 4 garçons est alors donnée par la loi binomiale $B_8^{1/2}$; c'est

$$p(\tfrac{1}{2}) = 70(\tfrac{1}{2})^8 \approx 0{,}2734$$

Le nombre X est alors une variable aléatoire de loi binomiale $B_n^{p(\frac{1}{2})}$

Sa moyenne est $np(\frac{1}{2})$, son écart-type $\sqrt{n.p(\frac{1}{2}).q(\frac{1}{2})}$,

où $q(\frac{1}{2}) = 1 - p(\frac{1}{2})$.

D'après ce qui précède et en utilisant le résultat élémentaire 1.4 (et non le résultat précis 1.1), si n est grand on a :

$$\text{Prob}\{|X - np(\tfrac{1}{2})| > 2{,}6\sqrt{np(\tfrac{1}{2})\,q(\tfrac{1}{2})}\} \leqslant 0{,}05$$

Donc a fortiori

$$\text{Prob}\{X > np(\tfrac{1}{2}) + 2{,}6\sqrt{np(\tfrac{1}{2})\,q(\tfrac{1}{2})}\} \leqslant 0{,}05$$

Dans la suite : $\alpha = np(\frac{1}{2}) + 2{,}6\sqrt{np(\frac{1}{2})\,q(\frac{1}{2})}$.

2.2 ***Hypothèse 2*** :

"Il existe un nombre $P \in [0;1]$ tel que lors de toute naissance il y ait probabilité P de naissance d'un garçon".

La probabilité que parmi les 8 enfants d'une famille il y ait exactement 4 garçons est alors donnée par la loi binomiale B_8^p ; c'est $p(P) = 70\,P^4(1-P)^4$; on observe que cette probabilité est plus faible que précédemment : $p(P) \leqslant p(\frac{1}{2})$. Il est dès lors parfaitement intuitif que la probabilité que $X > \alpha$ s'en trouve diminuée.

Exercice : Si Z suit la loi binomiale B_n^p et Z′ suit la loi binomiale $B_n^{p'}$ avec $p' < p$, pour tout λ on a :

$$\text{Prob}\{Z \geqslant \lambda\} \geqslant \text{Prob}\{Z' \geqslant \lambda\}$$

Par suite on a encore $\text{Prob}\{X > \alpha\} \leqslant 0{,}05$.

2.3 *Hypothèse 3* :

"Pour chaque couple ω, il existe un nombre $P(\omega) \in [0;1]$ tel que pour ce couple à chaque naissance la probabilité de naissance d'un garçon est $P(\omega)$ (insistons : $P(\omega)$ ne varie pas avec l'âge et n'est pas modifié par le sexe des enfants déjà nés)".

Pour chaque couple ω, il y a alors une probabilité a priori $p(P(\omega)) = 70\,P(\omega)^4\,(1-P(\omega))^4 \leqslant p(\frac{1}{2})$ d'avoir 4 garçons et 4 filles au cours de 8 naissances.

Le raisonnement rigoureux demande peut-être un peu plus d'habileté, mais la conclusion semble encore tout autant évidente :

$$\text{Prob}\{X > \alpha\} \leqslant 0{,}05 .$$

Conclusion : sous chacune des hypothèses 1,2,3, la probabilité que

$$(X > np(\tfrac{1}{2}) + 2{,}6\sqrt{np(\tfrac{1}{2})\,q(\tfrac{1}{2})})$$

est inférieure à 0,05 *avec* $p(\frac{1}{2}) = 70\,(\frac{1}{2})^8$ et $q(\frac{1}{2}) = 1 - p(\frac{1}{2})$.

En particulier, si $n = 53\,680$, on trouve

$$\text{Prob}\{X > 14\,947\} \leqslant 0{,}05$$

3. Test de l'hypothèse $\mathcal{H}$ au seuil 5%

3.1 *Construction du test*

$\mathcal{H}$ désigne l'une quelconque des hypothèses 1, 2, 3 envisagées. Il résulte du 2° que si $\mathcal{H}$ est vérifiée, alors, à plus de 95 chances sur 100, si nous choisissons au hasard 53 680 familles de 8 enfants, il y aura parmi ces familles un nombre de familles ayant 4 garçons et 4 filles au plus égal à 14 947.

D'où le principe du test :

i) Si nous trouvons un nombre de familles $\leq$ 14 947, on acceptera l'hypothèse $\mathcal{H}$ (ou plutôt on ne la rejettera pas) ; notre prédiction basée sur $\mathcal{H}$ s'est accomplie.

ii) Si nous trouvons un nombre de familles $>$ 14 947, on rejettera l'hypothèse $\mathcal{H}$. Il y a évidemment un risque couru : celui de rejeter à tort l'hypothèse $\mathcal{H}$ alors qu'elle est vraie, induits en erreur par un échantillon non représentatif ; ce risque, ou seuil, est (moindre que) 5%.
Précisons : supposons l'hypothèse $\mathcal{H}$ vérifiée, le résultat du test dépend du hasard de l'expérimentation, le rejet (à tort) de l'hypothèse $\mathcal{H}$ est un événement de probabilité $\leq$ 5%.

Note : l'intervention du "hasard" peut être comprise dans le choix des familles parmi les familles de 8 enfants, mais surtout plus fondamentalement dans le fait que la constitution de ces familles observées n'est qu'un état qui s'est réalisé parmi d'autres possibles. On voit qu'il y a là (et d'ailleurs aussi dans la formulation des hypothèses 1, 2 et 3) source de discussions pataphilosophiques sans fin qu'il vaut mieux éviter : Vive la notion d'espace probabilisé ! Un modèle mathématique, ici probabiliste, nous est proposé pour le problème des naissances ; il ne s'agit pour nous que de savoir si les calculs qui pourront y être développés seront en accord avec l'expérience. Une meilleure formulation des hypothèses 1, 2 et 3 serait donc : "Tout se passe comme si ...".

3.2 *Le test*. Maintenant et *maintenant seulement*, passons aux résultats expérimentaux.

Voici le résultat d'une enquête portant sur 53 680 familles. On a trouvé 14 959 familles de 4 garçons et 4 filles. Nous rejetons donc l'hypothèse $\mathcal{H}$ au seuil 5% !

Commentaires : Je ne suis ni probabiliste, ni statisticien, mais cela a dû se voir. Le rejet de l'hypothèse est basé sur un fait surprenant, pour moi : un excès de familles bien équilibrées à 4 filles et 4 garçons. Je dois enfin avouer n'avoir pas suivi les règles de l'art, et c'est une erreur grave. C'est après avoir considéré les résultats expérimentaux et noté l'excès de familles bien équilibrées que j'ai conçu le test. Un test véritable des hypothèses $\mathcal{H}$ selon la procédure exposée en 3.1 nécessiterait la donnée de nouveaux chiffres expérimentaux, les résultats exposés ici ne servant qu'à indiquer la voie d'un rejet éventuel de l'hypothèse $\mathcal{H}$.

Le seuil 5% a été choisi car il semble couramment utilisé.

Insistons bien sur un point important : il faut parfaitement définir le test (mécanisme d'acceptation ou de rejet de l'hypothèse) **avant** de considérer les résultats expérimentaux.

Ceci n'interdit pas toutefois, dans notre exemple, de décider le rejet de l'hypothèse si

$$X > 53\,680\, p(\tfrac{1}{2}) + k\sqrt{53\,680\, p(\tfrac{1}{2})\, q(\tfrac{1}{2})},$$

où k est tel que, si Y suit la loi binomiale $B_{53680}^{p(1/2)}$, la probabilité que

$$Y > 53\,680\, p(\tfrac{1}{2}) + k\sqrt{53\,680\, p(\tfrac{1}{2})\, q(\tfrac{1}{2})}$$

est $\leqslant 0{,}05$, mais d'attendre les résultats expérimentaux pour affiner l'estimation sur k jusqu'à obtenir éventuellement le rejet de l'hypothèse. Ainsi nous sommes allés jusqu'à la considération du moment d'ordre 6.

Note : les chiffres expérimentaux fournis dans le livre déjà cité de F. Gremy et D. Salmon sont donnés dans le tableau suivant, où, sur un total de 53 680 familles, N est le nombre de familles ayant k garçons :

k	0	1	2	3	4	5	6	7	8
N	215	1 485	5 331	10 649	14 959	11 929	6 678	2 092	342

Enfin, de l'hypothèse $\mathcal{H}$ on peut évidemment déduire d'autres conséquences que celle dégagée en 2° ; on peut donc bâtir d'autres tests de l'hypothèse. Le test ici présenté a l'avantage de se baser sur une théorie complètement élémentaire (comparer à la théorie d'un test du χ^2 !!) et de ne demander l'introduction d'aucune hypothèse supplémentaire. Par contre, si nous avions attendu un nombre de familles de 4 garçons et 4 filles "suffisamment inférieur" à $np(\frac{1}{2})$, on pourrait douter du bien-fondé de l'hypothèse $\mathcal{H}$, parce que nous savons bien par l'expérience que lors d'une naissance la probabilité d'avoir un garçon est voisine de 1/2 (heureusement !) ; en l'absence d'encadrement de cette probabilité, fourni par l'expérience, il ne nous est pas possible de préciser l'expression "suffisamment inférieur" (pour l'hypothèse 3, songer au cas où tout couple ne pourrait avoir que des garçons ou que des filles, i.e. $P(\omega) = 0$ ou 1).