Poolage

Le poolage est une méthode, assez controversée[1] du reste, utilisée pour tester des prélèvements sanguins et y détecter la présence d'un caractère rare. Au lieu de tester un par un les prélèvements provenant de plusieurs patients, on les mélange, formant ainsi une sorte de cocktail, qui donne lieu à un test unique.

La plupart des manuels de terminale proposent un exercice sur le thème « poolage », en relation généralement avec la loi binomiale, par exemple Déclic TES édition 1998, 57 page 252. Mais on doit trouver, j'imagine, dans des manuels plus récents.

Voici l'occasion de faire connaissance avec Pamela Borantine, héroïne d'exercices sur les tests de dépistage d'un manuel canadien dont j'ai malheureusement oublié les références. Elle figure dans un exercice frère jumeau de la partie A ci-dessous que l'on trouve sur le site TI-nSpire (http://www.univers-ti-nspire.fr/files/pdf/TI-Nspire_chap14_capes.pdf). Il y sera question de conseiller Pamela de façon à ce qu'elle puisse partir le plus tôt possible en week-end faire du ski dans les Montagnes Rocheuses.

1. Le sujet

Partie A : Pamela part en week-end

« Ce vendredi après midi à 14 heures, Pamela Borantine doit effectuer dans son laboratoire un test sur 16 échantillons pour détecter un caractère C. Elle sait par expérience que la probabilité qu'un échantillon donné possède le caractère C est 0,1. Les $_{gj}$ échantillons dont elle dispose sont indépendants.

Cependant, chaque test demande 15 minutes de travail, et Pamela craint de finir très tard son travail et de manquer son train Rocky Mountaineer de 18h32.

1. Une idée lui vient à l'esprit : si elle prélevait la moitié de chaque échantillon, et les groupait deux par deux, elle n'aurait plus que 8 tests à effectuer. Au cas où un groupe serait positif, elle n'aurait plus qu'à faire des tests séparément sur les moitiés restantes des échantillons. Elle pense qu'il n'y aura sûrement pas beaucoup de groupes $_{gj}$ positifs, et qu'elle a de fortes chances de gagner du temps.

A-t-elle raison ? À quelle heure peut-elle raisonnablement espérer finir son travail ?

2. Encouragée, Pamela se dit qu'elle gagnerait encore plus de temps si elle prélevait le quart de chaque échantillon et si elle formait 4 groupes de 4, quitte à analyser séparément les échantillons dont le groupe serait positif. A-t-elle raison ? »

[1] Une controverse est que les laboratoires pratiquant cette méthode facturent, dit-on, exhaustivement les analyses.

Partie B : Une généralisation

Il est question ici de dépistage d'une maladie concernant statistiquement une proportion connue p de la population, proportion que l'on suppose « faible » (on précisera), sur un ensemble de N personnes.
On considère que les résultats des tests individuels constituent des événements indépendants.

Dans la **méthode A**, on effectue N analyses individuelles (méthode exhaustive).

Dans la **méthode B**, on se donne un entier $r \geq 2$ et on répartit les N échantillons à tester par groupes de r échantillons. Dans chaque groupe, on prélève une part de chacun des r échantillons et on mélange les r prélèvements pour en faire une seule analyse, ce qui conduit à une première série de $\frac{N}{r}$ analyses.

(Implicitement, on suppose donc que N est un multiple de r. Si tel n'était pas le cas, on admet qu'un groupe puisse être incomplet et on néglige l'influence de ce groupe incomplet sur les résultats ; ceci a un sens dès que N est supposé « grand » par rapport à r, pour donner un ordre de grandeur dès que $\frac{N}{r} \geq 10$)

- Un groupe est négatif lorsque aucun des échantillons qui le composent n'est celui d'une personne contaminée.
- Sinon, le groupe est positif et l'on procède alors à une nouvelle analyse de chacun des r échantillons de ce groupe.

Étude de la méthode B

1. Quelle est la probabilité q_r pour qu'un groupe donné soit négatif ? En déduire la probabilité p_r pour qu'un groupe soit positif.

2. Soit X_r la variable aléatoire égale au nombre de groupes positifs.

2.1. Quelle est la loi de probabilité suivie par X_r ?

2.2. En déduire l'espérance $E(X_r)$ de la variable aléatoire X_r.

3. 1. Si X_r désigne le nombre de groupes positifs, calculer en fonction de X_r le nombre N_r d'analyses faites au total avec la méthode B.

3.2. Calculer l'espérance mathématique $E(N_r)$ de N_r.

3.3. Etablir que : gj $E(N_r) < N \Leftrightarrow 1 < r(1-p)^r$

Partie C : Comparaison d'un poolage avec la méthode exhaustive

Le but de cette partie est de préciser dans quelle mesure la méthode de poolage est plus économique que la méthode exhaustive et, si possible, de préciser comment optimiser cette méthode.

1. Dans cette question, on envisage des groupements deux par deux. Calculer l'espérance mathématique de N_2 (tel qu'il a été défini dans la partie B) et déterminer pour quelles valeurs de p l'inégalité $E(N_2) < N$ est vérifiée.

2. Une inégalité auxiliaire.

Soit r un nombre entier naturel au moins égal à 2. Démontrer que, pour tout réel x tel que $x \in [0\,;1]$:

$$1 - r\,x + \frac{r(r-1)}{2}x^2 \geq (1-x)^r \geq 1 - r\,x$$

3. Etablir que : $N\left(\dfrac{1}{r} + r\,p\right) \geq E(N_r) \geq N\left(\dfrac{1}{r} + r\,p - \dfrac{r(r-1)}{2}p^2\right)$

4. Montrer qu'une condition suffisante pour que $E(N_r) < N$ est que : $r^2\,p - r + 1 < 0$.

Le réel p étant connu et supposé être $\leq \dfrac{1}{4}$, comment choisir r pour que cette condition suffisante soit assurée ?

5. Le réel p est ici supposé connu et « faible », en l'occurrence $p \leq 0{,}1$. Pour optimiser la méthode de poolage, on se propose de choisir r de façon à minimiser non pas la fonction $r \mapsto E(N_r)$ elle-même mais sa fonction d'approximation $r \mapsto N\left(\dfrac{1}{r} + r\,p\right)$. En fonction de p, comment choisir r pour cela ?

Partie D : Application numérique

Cette fois, le laboratoire de Pamela doit tester un lot de 6000 échantillons pour y détecter un caractère C'. Pamela sait par expérience que la probabilité qu'un échantillon donné possède ce caractère C' est 0,005. Les échantillons sont indépendants.

Pamela envisage plusieurs façons d'organiser et d'appliquer le protocole décrit dans la « méthode B » :

- P*a*. Constituer 300 groupes de 20 prélèvements
- P*b*. Constituer 400 groupes de 15 prélèvements
- P*c*. Constituer 500 groupes de 12 prélèvements
- P*d*. Constituer 600 groupes de 10 prélèvements

Dans chaque cas, il s'agit de tester tous les groupes et, si le groupe est positif, chaque échantillon du groupe.

1. Quel choix lui conseiller ?

2. Elaborer une simulation permettant d'illustrer ces protocoles et de conjecturer quel est parmi eux le protocole le plus performant.

2. Eléments de correction

Partie A : Pamela part en week-end

Si elle effectue les 16 tests de façon exhaustive, Pamela mettra 240 minutes, soit 4 heures pour les effectuer et terminera à 18 heures. (Il lui restera 32 minutes pour chercher ses skis, aller à la gare et monter dans le train …)

1. Si elle les groupe deux par deux, la probabilité qu'un groupe donné soit négatif est $0,9^2 = 0,81$. La variable aléatoire X = « nombre de groupes positifs » suit la loi $B(8\,;0,19)$. Elle a pour espérance : 1,52

Le nombre N_1 de tests à effectuer, suivant ce procédé est égal à : $N_1 = 8 + 2X$ puisque chaque groupe positif nécessite deux tests supplémentaires. Par $_{gj}$ linéarité de l'espérance : $E(N_1) = 8 + 2\times 1,52 = 11,04$

Chaque test demandant 15 minutes de travail, la durée de travail est $D_1 = 15N_1$ et son espérance est 165,6 minutes. Pamela peut espérer finir à 16 h 45 min 36 s …

Elle peut espérer un gain de plus d'une heure par rapport à la méthode exhaustive.

2. Si Pamela groupe ses échantillons 4 par 4 : la probabilité qu'un groupe donné soit négatif est $0,9^4 = 0,6561$. La variable aléatoire X = « nombre de groupes positifs » suit la loi $B(4\,;0,3439)$. Elle a pour espérance : 1,3756

Le nombre N_2 de tests à effectuer, suivant ce procédé est égal à : $N_2 = 4 + 4X$ puisque chaque groupe positif nécessite quatre tests supplémentaires. Par linéarité de l'espérance : $E(N_2) = 4 + 4\times 1,3756 = 9,5024$

Chaque test demandant 15 minutes de travail, la durée de travail est $D_2 = 15N_2$ et son espérance est 142,536. Pamela peut espérer finir à 16 h 22 min 32 s …

Cette solution est plus performante que la précédente. On peut conseiller à Pamela de constituer quatre groupes de quatre. Elle peut espérer finir son travail plus tôt que dans les deux autres méthodes.

L'écran ci-contre propose la distribution de probabilité associée à la loi $B(8\,;0,19)$. On retrouve en cellule E2 l'espérance mathématique du temps d'analyse.

La cellule E1 affiche la probabilité que le temps d'analyse soit inférieur à celui de la méthode exhaustive. On note toutefois que Paméla court un risque, certes très minime, de voir son temps de travail augmenter et donc de manquer son train.

	A	B	C	D	E	F
=	=seq(k,k,0,8)	$.19)^{a[⬚]}\cdot(0.81)^{8-a[⬚]}$	=15*(8+2*a[])	=b[]*c[]		
1	0	0.185302	120	22.2362	0.952438	
2	1	0.347727	150	52.1591	165.6	
3	2	0.28548	180	51.3864	gilbertjulia2017	
4	3	0.133929	210	28.125		
5	4	0.039269	240	9.42462		
6	5	0.007369	270	1.98964		
7	6	0.000864	300	0.259281		
8	7	0.000058	330	0.019115		
9	8	0.000002	360	0.000611		
10						
11						
12						

B $=\mathrm{ncr}(8,a[⬚])\cdot(0.19)^{a[⬚]}\cdot(0.81)^{8-a[⬚]}$

L'écran ci-contre propose la distribution de probabilité associée à la loi $B(4\,;0,3439)$. On retrouve en cellule E2 l'espérance mathématique du temps d'analyse, inférieure à celle du cas précédent.

La cellule E1 affiche la probabilité que le temps d'analyse soit inférieur à celui de la méthode exhaustive. On note que Paméla court toutefois un risque un peu plus grand de manquer son train.

	A	B	C	D	E	F
=	=seq(k	$)^{a[⬚]}\cdot(0.6561)^{4-a[⬚]}$	=15*(4+4*a[])	=b[]*c[]		
1	0	0.185302	60	11.1181	0.879273	
2	1	0.38851	120	46.6212	142.536	
3	2	0.305461	180	54.983	gilbertjulia2017	
4	3	0.10674	240	25.6176		
5	4	0.013987	300	4.19614		
6						
7						
8						
9						
10						
11						
12						

B $=\mathrm{ncr}(4,a[⬚])\cdot(0.3439)^{a[⬚]}\cdot(0.6561)^{4-a[⬚]}$

Partie B : Une généralisation

1. $q_r = (1-p)^r$. Un groupe est positif si et seulement si au moins un échantillon de ce groupe est positif, évènement contraire de « le groupe est négatif » : $p_r = 1-(1-p)^r$.

2. X_r suit la loi binomiale $B\left(\frac{N}{r}\ ;\ p_r\right)$. En conséquence $E(X_r) = \frac{N}{r} \times p_r = \frac{N}{r}\left(1-(1-p)^r\right)$

3. On dénombre les $\frac{N}{r}$ analyses initiales et r analyses supplémentaires pour chaque groupe positif. Ainsi : $N_r = \frac{N}{r} + r\,X_r$

Par linéarité de l'espérance : $E(N_r) = \frac{N}{r} + r\,E(X_r) = N\left(\frac{1}{r} + \left(1-(1-p)^r\right)\right) =_{\text{gilbertjulia2017}} N + N\left(\frac{1}{r} - (1-p)^r\right)$

4. $E(N_r) < N \Leftrightarrow \frac{1}{r} - (1-p)^r < 0$.

Partie C

1. Si on effectue des groupement deux par deux :
Un groupe est positif si et seulement si au moins un échantillon de ce groupe est positif : $p_2 = 1-(1-p)^2 = 2p - p^2$. X_2 suit la loi binomiale $B\left(\frac{N}{2}\ ;\ 2p-p^2\right)$.

En conséquence $E(X_2) = \frac{N}{2} \times (2p - p^2) = N\left(p - \frac{p^2}{2}\right)$ et $E(N_2) = \frac{N}{2} + 2E(X_2) = N\left(\frac{1}{2} + 2p - p^2\right)$

$E(N_2) < N \Leftrightarrow \frac{1}{2} < (1-p)^2$ soit $E(N_2) < N \Leftrightarrow p < 1 - \frac{\sqrt{2}}{2}$

L'espérance mathématique du nombre d'analyses est plus petite que dans la méthode exhaustive lorsque $p < 1 - \frac{\sqrt{2}}{2}$, soit $p \le 0{,}29$ à 10^{-2} près. Un éventuel poolage n'a de sens que lorsque p vérifie cette inégalité.
On considèrera désormais que p est « faible » dès lors que p est plus petit que cette valeur.

2. Une inégalité auxiliaire.

On définit : $h_r(x) = 1-(1-x)^r - r\,x$: $h_r{}'(x) = r(1-x)^{r-1} - r = r\left((1-x)^{r-1}-1\right)$

Pour $x \in [0\,;1]$, h_r' est une fonction négative. La fonction h_r est décroissante sur $[0\,;1]$ et elle s'annule en zéro : elle est négative sur cet intervalle.

On définit aussi : $k_r(x) = 1-(1-x)^r - r\,x + \dfrac{r(r-1)}{2}x^2$:

$k_r{}'(x) = r(1-x)^{r-1} - r + r(r-1)x = r\left((1-x)^{r-1} - 1 + (r-1)x\right) = -r\,h_{r-1}(x)$

Pour $x \in [0\,;1]$, k_r' est une fonction positive. La fonction k_r est décroissante sur $[0\,;1]$ et elle s'annule en zéro : elle est positive sur cet intervalle.

Pour $x \in [0\,;1]$: $1-(1-x)^r - r\,x + \dfrac{r(r-1)}{2}x^2 \geq 0 \geq 1-(1-x)^r - r\,x$.

D'où on déduit dans ces mêmes conditions : $1 - r\,x + \dfrac{r(r-1)}{2}x^2 \geq (1-x)^r \geq 1 - r\,x$

3. On a vu que : $E(N_r) < N \Leftrightarrow \dfrac{1}{r} < (1-p)^r$.

Sachant, en vertu de l'inégalité auxiliaire, que : $(1-p)^r \geq 1 - r\,p$:

$1 - r\,p > \dfrac{1}{r} \Rightarrow E(N_r) < N$ (*a fortiori*).

$\dfrac{1}{r} - (1 - r\,p) = \dfrac{r^2 p - r + 1}{r}$

Une condition suffisante pour que $E(N_r) < N$ est que : $r^2\,p - r + 1 < 0$.

Le discriminant de l'expression $r^2\,p - r + 1$ est $1 - 4\,p$.

Si $p \geq \dfrac{1}{4}$, ce discriminant est négatif et l'expression $r^2\,p - r + 1$ est toujours positive. On n'obtient aucune condition suffisante assurant l'inégalité visée. On peut s'en douter vu la question C1.

Sinon, $r^2\,p - r + 1 < 0$ lorsque : $\dfrac{1-\sqrt{1-4p}}{2p} < r < \dfrac{1+\sqrt{1-4p}}{2p}$. Il s'agit là d'une condition suffisante pour que $E(N_r) < N$. On note que la condition $\dfrac{1-\sqrt{1-4p}}{2p} < r$ n'est pas vraiment utile car lorsque $0 < p < \dfrac{1}{4}$, $1 < \dfrac{1-\sqrt{1-4p}}{2p} < 2$ (ce qui signifie que le poolage est « rentable » à partir d'un mélange de 2 prélèvements au moins, ce que la question 1 a déjà examiné).

Il reste la condition suffisante : $2 \leq r < \dfrac{1+\sqrt{1-4p}}{2p}$

4. En appliquant la double inégalité : $N\left(\dfrac{1}{r} + r\,p\right) \geq E(N_r) \geq N\left(\dfrac{1}{r} + r\,p - \dfrac{r(r-1)}{2}p^2\right)$, on obtient que : $0 \leq E(N_r) - N\left(\dfrac{1}{r} + r\,p\right) \leq N \times \left(\dfrac{r(r-1)}{2}p^2\right)$ soit : $0 \leq E(N_r) - N\left(\dfrac{1}{r} + r\,p\right) \leq N \times \left(\dfrac{r(r-1)}{2}p^2\right)$ ce qui justifie le choix de $N\left(\dfrac{1}{r} + r\,p\right)$ comme fonction d'approximation de l'espérance pour « p petit ».

La fonction $r \mapsto \left(\frac{1}{r}+r\,p\right)$ est minimale pour $r=\frac{1}{\sqrt{p}}$.

Ainsi, l'optimisation envisagée consiste à choisir pour r l'entier « le plus proche de $\frac{1}{\sqrt{p}}$ ». L'espérance mathématique $E(N_r)$ est dans ce cas de l'ordre de $2N\sqrt{p}$

Attention, ce choix n'est pas forcément la valeur de r qui rend réellement minimale l'espérance. Déjà, il peut y avoir deux entiers candidats dès lors que $\frac{1}{p}$ n'est pas le carré d'un entier (l'entier juste avant et celui juste après $\frac{1}{\sqrt{p}}$). Ensuite, c'est la fonction $r \mapsto N\left(\frac{1}{r}+r\,p\right)$ qui a été minimisée et non l'espérance elle-même. Il se peut qu'il y ait un léger décalage entre les minimisations des deux fonctions.

Partie D.

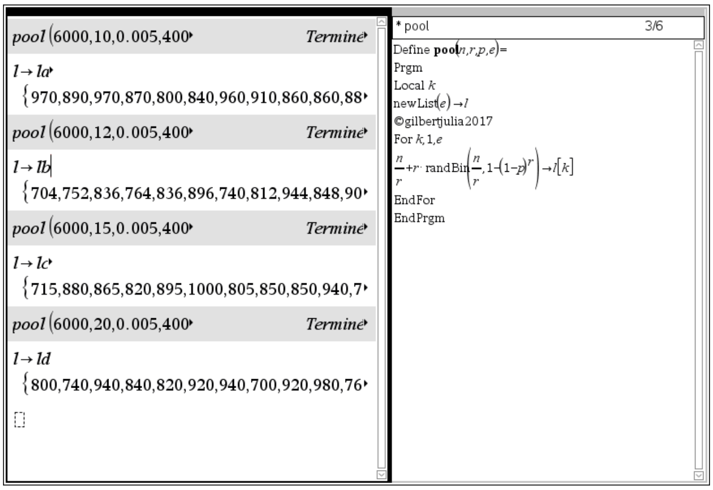
Si $p=0{,}005$: $\frac{1+\sqrt{1-4p}}{2p}=198{,}9$ à 0,1 près. Le poolage est à coup sûr rentable si $2 \le r \le 198$

De plus, le nombre réel $\frac{1}{\sqrt{0{,}005}}$ est entre 14 et 15. Le choix de r envisagé par Pamela le plus proche de ce nombre est l'entier 15.

On peut conseiller à Pamela de constituer 400 groupes de 15 prélèvements.

Une simulation permet de considérer ce qu'il se passe dans chacun des cas $r=10$; 12 ; 15 ; 20

On effectue une simulation de 400 tests suivant chacun des 4 protocoles.

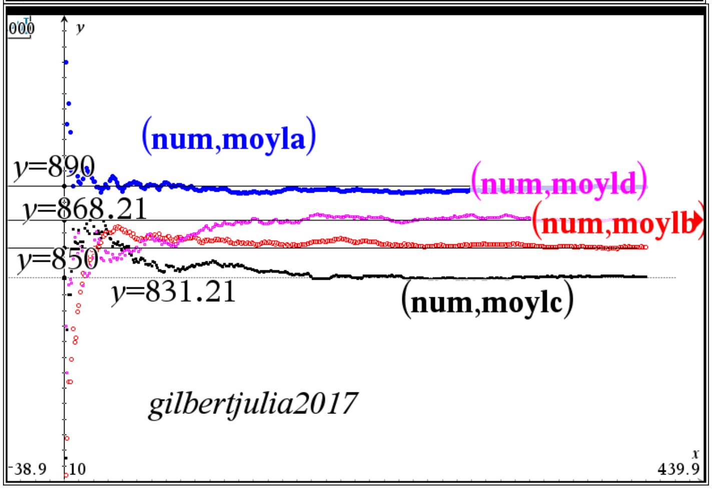


Les listes **moyla**, **moylb**, **moylc** et **moyld** désignent les nombres moyens d'analyses nécessaires dans chacun des 4 protocoles.

	A num	B moyla	C moylb	D moylc	E moyld	F	G
=	=seq(k	=cumulativesum(la)/num	=cumulati	=cumulati	=cumulati		
1	1.	970.	704.	715.	800.	gilbjulia2017	
2	2.	930.	728.	797.5	770.		
3	3.	943.333	764.	820.	826.667		
4	4.	925.	764.	820.	830.		
5	5.	900.	778.4	835.	828.		
6	6.	890.	798.	862.5	843.333		
7	7.	900.	789.714	854.286	857.143		
8	8.	901.25	792.5	853.75	837.5		
9	9.	896.667	809.333	853.333	846.667		
10	10.	893.	813.2	862.	860.		
11	11.	891.818	821.818	854.091	850.909		

B moyla:= $\frac{\text{cumulativesum(la)}}{\text{num}}$

Une étude graphique par exemple amène à la conjecture que le troisième protocole ($r = 15$) semble le meilleur.



Cet écran affiche des valeurs approchées de l'espérance $E(N_r)$ pour r = 10,12,15,20 ainsi que des valeurs approchées des deux fonctions utilisées dans le problème qui encadrent cette espérance.

Define $a(n,r,p)=\frac{n}{r}+n\cdot r\cdot p$ *Terminé*

Define $b(n,r,p)=\frac{n}{r}+n\cdot r\cdot p-\frac{n\cdot r\cdot(r-1)\cdot p^2}{2}$ *Terminé*

©gilbertjulia2017

$\{e(6000,10,0.005),e(6000,12,0.005),e(6000,15,0.005),e(6000,20,0.005)\}$
$\{893.339,850.263,834.586,872.337\}$

$\{a(6000,10,0.005),a(6000,12,0.005),a(6000,15,0.005),a(6000,20,0.005)\}$
$\{900.,860.,850.,900.\}$

$\{b(6000,10,0.005),b(6000,12,0.005),b(6000,15,0.005),b(6000,20,0.005)\}$
$\{893.25,850.1,834.25,871.5\}$

$\frac{e(428\cdot 14,14,0.005)\cdot 6000}{428\cdot 14}$ 835.191

L'écart entre $E(N_r)$ et son approximation $\frac{6000}{r}+30r$ grandit avec r.

La minimisation de $E(N_r)$ peut ne pas avoir lieu exactement pour le même entier que celle de $\frac{6000}{r}+30r$

Par curiosité, on teste le choix $r = 14$. Il apparaît que constituer 428 groupes de 14 personnes et un groupe incomplet ne serait pas plus performant.