Mots de longueur donnée à base de *P* lettres, et fonction génératrice

Considérons les mots de longueur *n* à base de *P* lettres, avec *n* entier positif.

1) Combien existe-t-il de tels mots ?

La première lettre du mot est l'une des P lettres, soit P cas. A chaque fois il en est de même pour la deuxième lettre, et ainsi de suite. Le nombre des mots est P^n.

2) Combien existe-t-il de mots de longueur *n* ne contenant qu'une seule lettre ? Combien s'ils n'ont que deux lettres différentes ?

Si les mots sont tous formés avec la même lettre répétée, il y a P mots, autant qu'il y a de lettres possibles.

Si les mots de longueur n contiennent seulement deux lettres différentes, ce qui impose $n \geq 2$, on commence par choisir ces deux lettres, soit C_P^2 cas. Une fois les deux lettres choisies, il s'agit de former tous les mots ne contenant qu'elles. Il existe 2^n mots à base de deux lettres, mais parmi eux il faut enlever ceux qui ne contiennent qu'une seule des deux lettres, soit deux mots. Finalement le nombre de mots cherchés est $C_P^{\,2}(2^n - 2)$.

3) On suppose dans cette question que *n* = *P*. Combien existe-t-il de mots contenant toutes les lettres ? Combien en existe-t-il contenant exactement *K* des *P* lettres ?

Les mots de longueur P contenant les P lettres ne sont autres que les permutations de P éléments, leur nombre est $P!$.

Supposons maintenant qu'ils contiennent K lettres, avec K entre 1 et P. Commençons par choisir K lettres parmi P, soit C_P^K cas. Puis plaçons-nous dans un de ces cas. Pour connaître le nombre des mots ayant K lettres données, plaçons-nous dans le contexte de la formule du crible, en considérant K propriétés, la propriété i , avec i entre 1 et K, signifiant que la lettre i n'est pas présente dans le mot. Le nombre des mots cherchés est :

$$S_0 - S_1 + S_2 - \dots + (-1)^K S_K$$

(nous utilisons les notations du livre *Combien 1-Combinatoire, chapitre sur la formule du crible*).

Ici $S_0 = K^P$, $S_1 = C_K^1(K - 1)^P$, $S_2 = C_K^2(K - 2)^P$, …, $S_{K-1} = C_K^{K-1}(K - (K-1))^P$, $S_K = 0$.
D'où le nombre de mots :

$$C_P^K \left(K^P - C_K^1(K - 1)^P + C_K^2(K - 2)^P - \dots + (-1)^{K-1} C_K^{K-1}(K - (K - 1))^P \right)$$

Par exemple pour $P = 4$, le nombre des mots est $4^4 = 256$, et ils se divisent en quatre catégories:

- ceux contenant une seule des 4 lettres, soit 4 mots

- ceux contenant exactement deux lettres, soit $C_4^2\,(2^4 - 2) = 84$
- ceux contenant exactement trois lettres, soit $C_4^3\,(3^4 - 3.2^4 + 3) = 144$
- ceux contenant les quatre lettres, soit $C_4^4\,(4^4 - 4.\,3^4 + 6.\,2^4 - 4) = 24$

Le nombre moyen de lettres présentes dans ces mots est 2,7.

Remarquons au passage la formule : $\sum_{j=0}^{P} (-1)^j C_P^j (P-j)^P = P!$

4) On suppose dorénavant que les mots sont fabriqués au hasard. Cela revient à prendre une urne contenant *P* boules numérotées de 1 à *P*, et à effectuer *n* tirages successifs avec remise de la boule à chaque fois, ces tirages étant équiprobables. En notant les numéros des boules successivement tirées, on retrouve les mots de longueur *n* à base de *P* lettres. On appelle *Y*(*n*) la variable aléatoire correspondant au nombre de lettres (numéros) présentes dans ces mots de longueur *n*.

a) Calculer les probabilités *p*(*Y*(*n*) = 1) et *p*(*Y*(*n*) = 2).

Chaque mot a autant de chances d'être obtenu qu'un autre. Dans ce contexte d'équiprobabilité, la probabilité est le nombre de cas favorables sur le nombre de cas possibles. En utilisant les résultats du 1° et du 2°, on trouve :

$p(Y(n) = 1) = \dfrac{P}{P^n} = \dfrac{1}{P^{n-1}}$

$p(Y(n) = 2\,) = \dfrac{C_P^2(2^n - 2)}{P^n}$

b) Calculer *p*(*Y*(*n*)) = *n*) lorsque *n* est inférieur ou égal à *P*.

Les cas favorables sont constitués par les mots de longueur n ayant n lettres différentes. Cela revient à prendre les arrangements de *n* lettres prises parmi *P*, soit $A_P^{\,n}$. D'où la probabilité :

$p(Y(n)) = n) = \dfrac{A_P^n}{P^n}$

c) Etablir une relation de récurrence entre le nombre de mots *M*(*n* + 1, *k*) de longueur *n*+1 et ayant exactement k lettres présentes, et les nombres *M*(*n*, *k*) et *M*(*n*, *k* – 1) de mots de longueur *n* et ayant respectivement *k* et *k* – 1 lettres présentes.

Les $M(n + 1, k)$ mots de divisent en deux catégories, selon leur dernière lettre, la n+1 ème..

- Si la dernière lettre est déjà présente dans le reste du mot qui est de longueur *n* (avec *k* lettres présentes), il y a *k* dernières lettres possibles. On trouve ainsi $M(n, k)\ k$ mots.
- Si la dernière lettre n'est pas déjà présente dans le reste du mot de longueur *n* et qui contient $k - 1$ lettres différentes, il reste à choisir pour la dernière lettre une lettre autre que les $k - 1$, soit $P - k + 1$ cas. D'où le nombre de mots $M(n, k - 1)\,(P - k + 1)$.

On trouve finalement la relation :

$M(n+1, k) = k\,M(n, k) + (P-k+1)M(n, k-1)$

d) En déduire une relation de récurrence entre probabilités

On a, grâce à la relation précédente :

$$p(Y(n+1)=k) = \frac{M(n+1,k)}{P^{n+1}} = \frac{kM(n,k)+(P-k+1)M(n,k-1)}{P^{n+1}}$$

$$= \frac{k}{P}\,p(Y(n)=k) + \frac{P-k+1}{P}\,p(Y(n)=k-1)$$

5) On prend la fonction génératrice $G_n(X)$ associée à la loi de probabilité de Y_n, soit $G_n(X)=\sum_{k=1}^{+\infty} p(Y(n)=k)X^k$, mais comme cette somme infinie se termine par des 0, il suffit de prendre le polynôme en X :

$G_n(X)=\sum_{k=1}^{P} p(Y(n)=k)X^k$ (quelques-uns des termes de plus haut degré peuvent aussi être nuls lorsque $n < P$). En utilisant la relation de récurrence précédente, déterminer $G_{n+1}(X)$ en fonction de $G_n(X)$ et de sa dérivée $G'_n(X)$.

Constatons d'abord que $G'_n(X)=\sum_{k=1}^{P} k\,p(Y(n)=k)X^{k-1}$. En appliquant la formule du 4-d :

$$G_{n+1}(X)=\sum_{k=1}^{P} p(Y(n+1)=k)X^k$$

$$=\sum_{k=1}^{P}\frac{k}{P}\,p(Y(n)=k)X^k + \sum_{k=1}^{P}\frac{P-(k-1)}{P}\,p(Y(n)=k-1)X^k$$

$$=\frac{X}{P}\sum_{k=1}^{P} k\,p(Y(n)=k)X^{k-1} + \sum_{k'=0}^{P-1}\frac{P-k'}{P}\,p(Y(n)=k')X^{k'+1}$$

Dans le premier Σ, on voit apparaître la dérivée, et dans le deuxième Σ on a procédé à un changement de variable

$$=\frac{X}{P}G'_n(X)+\sum_{k'=1}^{P}\frac{P-k'}{P}p(Y(n)=k')X^{k'+1} \quad k' = k-1.$$

Dans le Σ, on a enlevé le terme avec $k' = 0$, puisqu'il est nul, et l'on a rajouté le terme avec $k' = P$, qui est nul lui aussi. Si l'on fait cela, c'est dans le but de voir apparaître les fonctions génératrices, comme cela va se préciser maintenant, en cassant le Σ en deux.

$$=\frac{X}{P}G'_n(X)+X\sum_{k'=1}^{P} p(Y(n)=k')X^{k'} - \frac{X^2}{P}\sum_{k'=1}^{P} k'p(Y(n)=k')X^{k'-1}$$

$$=\frac{X}{P}G'_n(X)+X\,G_n(X)-\frac{X^2}{P}G'_n(X)$$

$$G_{n+1}(X)=\frac{X-X^2}{P}G'_n(X)+X\,G_n(X)$$

6) Appelons $E(Y(n))$ la valeur moyenne (ou espérance) de la variable aléatoire $Y(n)$. Utiliser la formule précédente pour trouver une relation de récurrence entre $E(Y(n+1))$ et $E(Y(n))$. Puis donner la formule explicite de $E(Y(n))$.

On sait que $E(Y(n)) = G'_n(1)$. Comme on a aussi besoin de $E(Y(n+1)) = G'_{n+1}(1)$, on est amené à dériver la formule précédente.

$$G'_{n+1}(X) = \frac{X - X^2}{P} G''_n(X) + \frac{1-2X}{P} G'_n(X) + X\,G'_n(X) + G_n(X)$$

Faisons maintenant $X = 1$. La dérivée seconde a le bon goût de disparaître, et on sait aussi que $G_n(1) = 1$. Il reste :

$$E(Y(n+1)) = -\frac{1}{P} E(Y(n)) + E(Y(n)) + 1$$

$$= \frac{P-1}{P} E(Y(n)) + 1$$

On reconnaît la relation de récurrence d'une suite arithmético-géométrique, de la forme $u_{n+1} = q\, u_n + 1$, avec $q = \frac{P-1}{P}$. Son point fixe est égal à P. Introduisons la suite (v_n) telle que $v_n = u_n - P$. Par soustraction, on trouve :

$u_{n+1} = q\, u_n + 1$

$\underline{P \quad = q\, P + 1}$

$v_{n+1} = q\, v_n$ La suite (v_n) est géométrique, sa forme explicite est $v_n = q^{n-1} v_1$ avec $v_1 = u_1 - P$. Ainsi $u_n - P = q^{n-1}(u_1 - P)$ ou $u_n = q^{n-1}(u_1 - P) + P$. Avec ici $u_1 = E(Y(1)) = 1$, l'espérance est égale à :

$$E(Y(n)) = \left(\frac{P-1}{P}\right)^{n-1}(1-P) + P = P - \frac{(P-1)^n}{P^{n-1}} = P\left(1 - \left(\frac{P-1}{P}\right)^n\right)$$

7) On prend ici $n = P$. Faire le programme qui permet d'avoir $E(Y(P))$. Pour cela on réalise un grand nombre *NBE* d'expériences. Chaque expérience consiste à faire P tirages répétés avec remise dans une urne qui contient P boules numérotées de 1 à P. A chaque fois, on compte le nombre de numéros différents obtenus. En cumulant ces résultats de chaque expérience, et en divisant ce cumul par le nombre *NBE* d'expériences, on trouve la moyenne demandée $E(Y(P))$. Constater expérimentalement que le rapport $E(Y(P)) / P$ décroît lorsque P augmente, en ayant tendance à se stabiliser sur une valeur limite que l'on précisera.

Au cours de chaque expérience de P tirages, on utilise un tableau *fait*[] mis initialement à 0. Lorsqu'un numéro i est tiré, on met la case *fait*[*i*] à 1. Lorsque les P tirages sont effectués, la somme des 1 de ce tableau donne le nombre de numéros différents. On trouvera ci-dessous la partie stratégique du programme.

On se donne le nombre d'expériences NBE*, ainsi que la valeur de* P. *Comme les numéros vont de 1 à* P*, le tableau* fait*[] doit être déclaré* int fait*[*P+*1], la case 0 étant inoccupée.*

```
cumul=0;
for(experience=0;experience<NBE;experience++)
  {  for(i=1;i<=P;i++) fait[i]=0;
```

```
      for(tirage=1; tirage<=P; tirage++)
        { numero=1+random(P);  fait[numero]=1; }
      compteur=0;
      for(i=1;i<=P;i++) if (fait[i]==1) compteur++;
      cumul+=(float)compteur;
    }
  moyenne=cumul/(float)NBE;
  printf("\nN= %3.d valeur moyenne/P=: %3.4f ",P,moyenne/(float)P);
```

Le rapport $E(Y(P)) / P$ passe de 0,65 pour P = 10, à 0 ,634 pour P = 100, et 0,6324 pour P = 500. Il se stabilise sur une valeur de cet ordre.

8) Calculer la valeur limite de $E(Y(P)) / P$ en utilisant le résultat du 6°.

On sait que $E(Y(P)) = P\,(1-\left(\frac{P-1}{P}\right)^P)$ ou $\frac{E(Y(P))}{P}=(1-\left(\frac{P-1}{P}\right)^P)$.

Or $\left(\frac{P-1}{P}\right)^P = e^{P\ln(1-\frac{1}{P})}$ et l'on sait que pour P grand : $\ln(1-\frac{1}{P}) \approx -\frac{1}{P}$. Ainsi $\left(\frac{P-1}{P}\right)^P$ tend vers $\frac{1}{e}$. Le rapport $\frac{E(Y(P))}{P}$ tend vers $1 - \frac{1}{e}$, soit 0,6321.

Remarque finale : On savait que l'utilisation de fonctions génératrices était particulièrement efficace dans les problèmes de temps d'attente. On a là un exemple différent, où à défaut d'être connue explicitement, la fonction génératrice permet de calculer l'espérance.