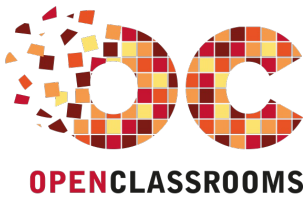
Les arbres de Galton-Watson

Par Mickaël Launay (GéoMI17)



www.openclassrooms.com

Dernière mise à jour le 20/11/2011

Sommaire

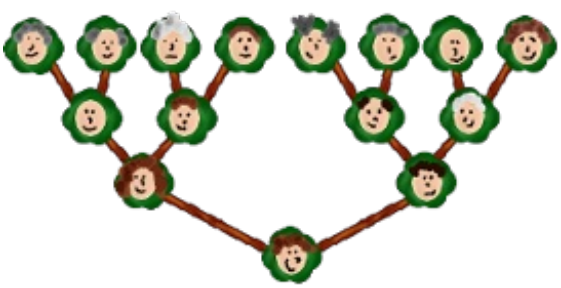
Les arbres de Galton-Watson

Par  Mickaël Launay (GéoMl17)

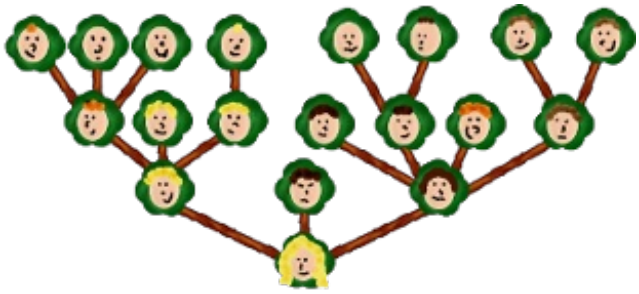
Mise à jour : 20/11/2011
Difficulté : Intermédiaire Durée d'étude : 5 heures

Peut-être vous êtes-vous déjà amusé à reconstituer votre arbre généalogique sur lequel vous avez reporté les noms de vos ancêtres, proches ou lointains. Mais vous êtes-vous déjà demandé à quoi allait ressembler **l'arbre de votre descendance** ?

L'arbre des ancêtres est très régulier : à chaque génération chaque branche se divise en deux. Autrement dit le nombre d'ancêtres double à chaque génération remontée. (Du moins à condition qu'il n'y ait pas de mariages entre cousins éloignés ce qui fini inévitablement par arriver si on remonte suffisamment loin.) Un arbre de descendance peut en revanche être beaucoup plus irrégulier car si chacun a deux parents, le nombre d'enfants peut quant à lui beaucoup varier !



Arbre des ancêtres



Arbre des descendants

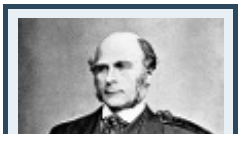
Nous allons voir qu'il est possible de décrire mathématiquement ces arbres de descendance. Comme il n'est bien entendu pas possible de prédire l'avenir, ce modèle est un modèle **probabiliste**, c'est-à-dire qu'il utilise la théorie des probabilités pour évaluer les chances des différents scénarios possibles. Ces arbres aléatoires sont appelés **les arbres de Galton-Watson**.

Quelques prérequis sont conseillés pour la lecture de ce cours. Tout d'abord il est recommandé d'avoir déjà quelques bases en probabilités ; même si tout est expliqué de zéro, le modèle n'est à mon avis pas le plus simple pour débuter en probas. Ensuite il est préférable de savoir étudier une fonction (dérivée, sens de variation,...) Si ce n'est pas le cas vous pouvez malgré tout suivre ce cours mais vous serez obligé de me faire confiance au moment où ce sera nécessaire. Enfin, à quelques endroits je détaillerais quelques calculs rigoureux qui s'adressent à des étudiants post-bac ayant déjà de bonnes bases en théorie des probabilités. Cependant, ces calculs seront toujours accompagnés d'explications intuitives pour que tout le monde puisse suivre même en sautant les calculs !

Sommaire du tutoriel :

Le modèle

L'histoire des arbres de Galton-Watson commence en 1873 lorsque le scientifique britannique Francis Galton se pose la question de l'évolution des noms de famille des lords anglais. Il s'inquiète de voir certains de ces noms disparaître après que leurs derniers représentants sont morts sans laisser de descendance. Il décide alors de poser sa question dans le journal *Educational Times* et reçoit peu de temps après une réponse avec la solution du révérend Henry William Watson.




Francis Galton

Depuis, les arbres de Galton-Watson (on a donné au modèle le nom de ses deux inventeurs) sont devenus des objets classiques de la théorie des probabilités et ont été étudiés en long en large et en travers avec de nombreuses variantes.

Dans ce mini-cours je vous propose simplement d'étudier le modèle de base de Galton et Watson et de voir comment il est possible de répondre à la question de départ : quelle est la probabilité pour que les noms des lords anglais ne s'éteignent pas. Ou plus généralement la probabilité pour que la descendance d'un individu (vous par exemple) ne s'éteigne pas.

La loi de natalité aléatoire

Bien, il est temps de rentrer dans le vif du sujet : comment définit-on un arbre de Galton-Watson ?

Avant de construire des arbres sur plusieurs générations, nous allons nous concentrer sur le nombre d'enfants d'un seul individu. Évidemment, tout le monde n'a pas le même nombre d'enfants, c'est donc à ce stade qu'il faut faire intervenir les probabilités.

On notera donc p_n la probabilité pour un individu d'avoir n enfants. Ici, n est un entier naturel : il n'est pas possible d'avoir un nombre négatif ou un nombre à virgule d'enfants ! Par contre, n peut-être égal à 0 : il est possible de ne pas avoir d'enfants.

En bref, on a une suite de probabilités :

- p_0 est la probabilité de ne pas avoir d'enfants ;
- p_1 est la probabilité d'avoir un seul enfant ;
- p_2 est la probabilité d'avoir deux enfants ;
- p_3 est la probabilité d'avoir trois enfants ;
- *et cætera*.

Si par exemple, on regarde les statistiques des années 1960 en France, on obtient les probabilités suivantes :

$$p_0 = 0,10, \quad p_1 = 0,18, \quad p_2 = 0,40, \quad p_3 = 0,22, \quad p_4 = 0,07, \quad p_5 = 0,03.$$

Ce qui signifie que 10% des gens n'ont pas d'enfants, 18% en ont un, 40% en ont deux, 22% en ont trois, 7% en ont quatre et 3% en ont cinq.

Et alors personne n'avait six enfants ou plus dans les années 60 ?

En réalité si. Mais il s'agit d'une minorité et pour plus de simplicité comme je vais me servir de cet exemple dans tout le cours j'ai préféré arrondir. (Comment ça fainéant ?)

Remarquez que la somme de ces probabilités doit être égale à 1 :

$$\sum_{n=0}^{+\infty} p_n = 1.$$

Si vous ne connaissez pas le signe $\sum$, il signifie simplement que la somme de tous les p_n est égale à 1. Si vous voulez en savoir plus sur cette notation et la façon dont on l'utilise vous pouvez lire ce chapitre de mon cours *Nombres et opérations*.

Le fait que cette somme soit égale à 1 signifie qu'un individu a forcément un nombre entier naturel d'enfants (0, 1, 2, 3, 4,...)

Comme je l'ai déjà dit, il n'est pas possible d'avoir un nombre négatif ou à virgule d'enfants.

Si on reprend l'exemple des années 60, on a bien : 0,10+0,18+0,40+0,22+0,07+0,03 =1.

Vous remarquerez dans la formule ci-dessus que la somme des p_n va de 0 jusqu'à... l'infini. Cela signifie que pour ce modèle on a pas besoin de limiter le nombre d'enfants par personne et qu'il est tout à fait possible que p_n soit différent de 0 pour n'importe quel nombre n.

Par exemple, on peut très bien avoir pour tout n :

$$p_n = \frac{1}{2^{n+1}},$$

c'est-à-dire $p_0 = 1/2, p_1 = 1/4, p_2 = 1/8, p_3 = 1/16$ et ainsi de suite chaque terme étant égal à la moitié du précédent. Cette suite vérifie bien l'égalité :

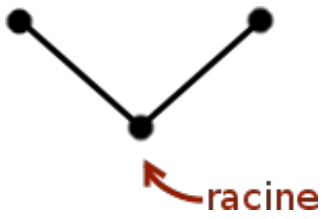
$$\sum_{n=0}^{+\infty} p_n = \sum_{n=0}^{+\infty} \frac{1}{2^{n+1}} = 1.$$

Évidemment dans l'exemple des années 60 ou dans toute autre situation concrète ça ne sert à rien de faire la somme jusqu'à l'infini puisque les p_n sont nuls à partir d'un certain rang (à partir de 6 dans l'exemple). Cependant comme les mathématiciens n'aiment pas faire les choses à moitié, ils préfèrent considérer le cas général où le nombre d'enfants peut-être aussi grand qu'on veut. Vous allez voir que ça ne change strictement rien pour l'étude théorique du modèle.

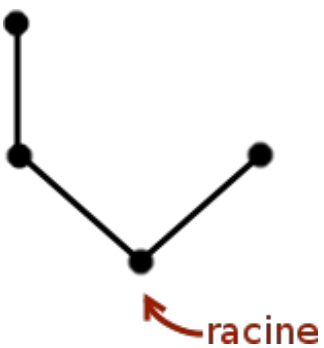
Les arbres

Maintenant que nous avons défini la loi de natalité, nous pouvons enfin construire nos arbres aléatoires. Pour cela, commençons par poser l'ancêtre commun de notre généalogie, que l'on appelle la racine de l'arbre :

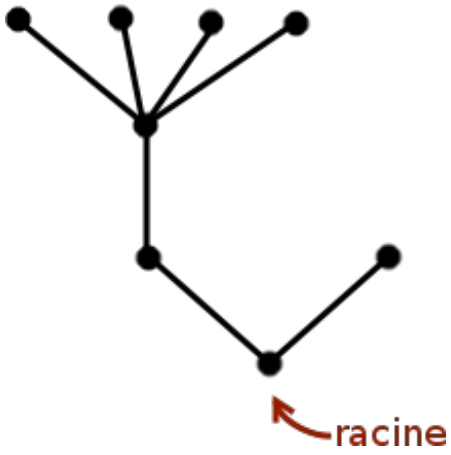
Cet ancêtre va alors engendrer un nombre aléatoire d'enfants. Par exemple, il y a une probabilité p_2 pour qu'il ait deux enfants :



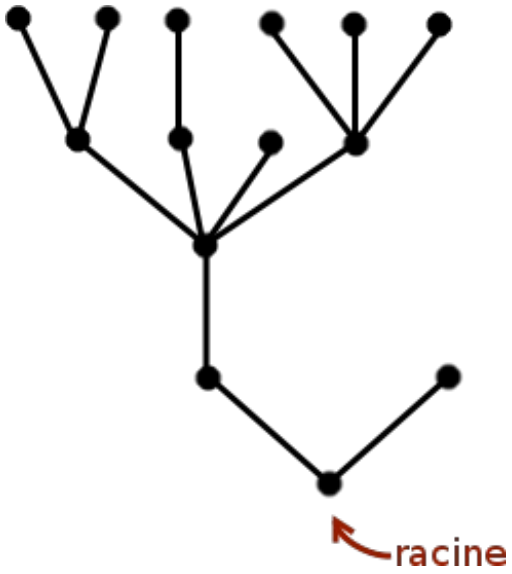
Puis chacun de ces deux enfants va à son tour avoir des enfants indépendamment. Par exemple, il y a une probabilité p_1 pour que le premier ait un enfant et une probabilité p_0 pour que le deuxième n'en ait pas. Il y a donc au total une probabilité $p_0 \times p_1$ pour que la deuxième génération de l'arbre soit la suivante :



Ensuite on continue sur le même principe. Il y a une probabilité p_4 pour que le seul individu de la deuxième génération ait quatre enfants :



Normalement vous devez commencer à comprendre le principe. Allez, une dernière génération. Il y a une probabilité $p_2 p_1 p_0 p_3$ pour que les quatre individus de la troisième génération aient respectivement 2, 1, 0 et 3 enfants.



Bon, on s'arrête ici, mais bien sûr on pourrait continuer ainsi à construire les différentes générations successivement en calculant leurs probabilités.

Si on récapitule tout, la probabilité que les quatre premières générations de l'arbre de Galton-Watson soient celles de l'exemple ci-dessus est égale à

$$(p_2) \times (p_1 p_0) \times (p_4) \times (p_2 p_1 p_0 p_3) = p_0^2 p_1^2 p_2^2 p_3 p_4 .$$

Et si on calcule cette valeur dans le cas particulier des familles de 1960 donné ci-dessus, on trouve que cet arbre a une probabilité égale à environ 0,0000008 ! Évidemment, c'est minuscule : cela fait environ une chance sur 1250000. Mais ce qu'il faut se dire c'est qu'il y a **énormément** de scénarios possibles. Il est donc normal que chacun d'entre eux soit peu probable.

Sur quatre générations, il n'est pas possible de trouver un arbre ayant une grande probabilité, ne serait-ce que de 1%. Et c'est encore pire sur 5, 6 ou davantage de générations...

Mais alors il est nul ton modèle ! Si l'on ne peut pas savoir ce qui se passe à plus d'une chance sur un million, c'est comme si on ne savait rien du tout de ce qui allait se passer ! Donc en fait on ne peut rien savoir de probable sur les arbres de Galton-Watson ?

Mais si, rassurez-vous, nous allons avoir des résultats intéressants ! Le problème vient du fait que nous ne nous sommes pas posé les bonnes questions.

Se demander si tel ou tel scénario précis va se produire est une question beaucoup trop pointue. D'autant que concrètement, ce n'est pas vraiment ce qui nous intéresse.

En revanche, il est beaucoup plus pertinent de se poser des questions qui sont à la fois plus larges et qui ont plus de sens comme *« Quelle est la probabilité d'avoir encore des descendants dans 1000 générations ? »*, *« Quelle est la probabilité d'avoir au moins trois petits-enfants ? »*, ou encore la question originelle de Galton *« Quelle est la probabilité pour que ma descendance s'éteigne à un moment donné ? »*

Ces questions-là ne demandent pas la probabilité d'un scénario précis, mais la probabilité d'un ensemble de scénarios qui vérifient une propriété donnée. Autrement dit, il s'agit de chercher la somme des probabilités de tous les arbres qui ont cette propriété.

Si on le dit de cette façon, la tâche peut paraître fastidieuse : trouver tous les arbres qui vérifient la propriété voulue, puis calculer leurs probabilités et les additionner. Rassurez-vous, ce n'est pas comme cela que l'on va procéder ! Il existe des méthodes bien plus élégantes et efficaces, comme nous allons le voir dans la deuxième partie de ce cours.

L'étude du modèle

Bien, alors maintenant que nous avons posé le modèle, nous allons pouvoir rentrer dans le vif du sujet : l'étude des propriétés générales de l'arbre. Autrement dit, c'est à partir de maintenant qu'on va commencer à faire des calculs.

Nombre moyen d'enfants

On sait que le nombre d'enfants d'un individu est aléatoire, oui mais combien en a-t-il *en moyenne* ?

Voilà une bonne question ! Si on ne peut pas donner précisément le nombre d'enfants, on peut toujours en donner la moyenne, ou en vocabulaire probabiliste, **l'espérance** qui se note avec la lettre $\mathbb{E}$.

Si vous êtes déjà un habitué des probabilités, cette question ne doit pas vous faire peur et la réponse est immédiate :

$$\mathbb{E}\left[\text{nombre d'enfants d'un individu}\right] = \sum_{n=0}^{\infty} n p_n.$$

Si au contraire cette formule n'a rien d'évident pour vous, voyons pourquoi elle est vraie en nous penchant sur l'exemple des années 60. Prenons un échantillon de 100 personnes. D'après les probabilités données au début de ce cours, on sait qu'en moyenne sur ces 100 personnes :

- 10 auront 0 enfant ;
- 18 auront 1 enfant ;
- 40 auront 2 enfants ;
- 22 auront 3 enfants ;
- 7 auront 4 enfants ;
- 3 auront 5 enfants.

Par conséquent, le nombre moyen d'enfants de ces 100 personnes réunies est égal à

$$0 \times 10 + 1 \times 18 + 2 \times 40 + 3 \times 22 + 4 \times 7 + 5 \times 3.$$

Et pour obtenir le nombre moyen d'enfants d'un seul individu, il reste à diviser ce nombre par 100. On trouve donc :

$$\frac{0\times 10+1\times 18+2\times 40+3\times 22+4\times 7+5\times 3}{100}=0\times\frac{10}{100}+1\times\frac{18}{100}+2\times\frac{40}{100}+3\times\frac{22}{100}+4\times\frac{7}{100}+5\times\frac{3}{100}.$$

On remarque alors que les fractions correspondent aux probabilités p_0, p_1, p_2, p_3, p_4 et p_5. Le nombre moyen d'enfants d'un individu est donc égal à

$$0\times p_0+1\times p_1+2\times p_2+3\times p_3+4\times p_4+5\times p_5.$$

Cette formule s'arrête à $5\times p_5$, car dans l'exemple le nombre d'enfants ne peut pas être supérieur à 5 mais vous comprenez que dans le cas général, il faut prolonger la formule par $6\times p_6$, $7\times p_7$, *et cætera*. La formule générale est donc bien celle annoncée :

$$\mathbb{E}\left[\text{nombre d'enfants d'un individu}\right]=\sum_{n=0}^{\infty}np_n.$$

Comme ce nombre va nous être très utile par la suite nous allons lui donner un nom : m. On pose donc

$$m:=\sum_{n=0}^{\infty}np_n.$$

Si on finit le calcul ci-dessus pour l'exemple des années 60, on trouve $m=2,07$.

On passe aux générations futures...

Que se passe-t-il pour les générations suivantes ? Combien en moyenne un individu a-t-il de petits-enfants, d'arrière-petits-enfants, d'arrière-arrière-petits-enfants, ...?

De façon intuitive on peut répondre à cette question de la manière suivante : un individu a m enfants en moyenne et chacun de ces enfants a à son tour m enfants. L'individu de départ a donc en moyenne $m\times m=m^2$ petits-enfants.

Puis chacun de ces m^2 petits-enfants va avoir en moyenne m enfants, d'où on déduit que notre individu va avoir m^3 arrière-petits-enfants. Et ainsi de suite, on comprend qu'en raisonnant de la sorte, l'ancêtre aura en moyenne m^k descendants à la $k^{\text{ème}}$ génération.

Le raisonnement que nous venons de faire est tout à fait correct. Cependant en mathématiques il est toujours préférable de vérifier par un calcul rigoureux (d'autant que la théorie des probabilités regorge de pièges qui semblent à première vue contraires à l'intuition et au bon sens).

Le calcul suivant s'adresse plutôt à des étudiants post-bac qui ont déjà de bonnes bases en théorie des probabilités. Si vous n'êtes pas dans ce cas et que le raisonnement intuitif ci-dessus vous a convaincu, vous pouvez sauter les quelques lignes suivantes.

Nous allons procéder par récurrence. Supposons donc que l'espérance du nombre de descendants à la génération k d'un individu soit m^k. Autrement dit, si on note N le nombre (aléatoire) de descendants à la génération k on a :

$$\mathbb{E}\left[N\right]=m^k.\ \text{(Hypothèse de récurrence)}$$

Et pour $1\leq i\leq N$ notons M_i le nombre (aléatoire lui aussi) d'enfants du $i^{\text{ème}}$ descendant de la génération k. Le nombre de descendants à la génération $k+1$ est alors :

$$\sum_{i=1}^{N} M_i.$$

Calculons donc l'espérance de ce nombre :

$$\mathbb{E}\left[\sum_{i=1}^{N} M_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} M_i \middle| N\right]\right]$$

On commence par écrire notre espérance en fonction de l'espérance conditionnellement à N. Notez que cette égalité n'a rien de compliqué, c'est juste la définition de l'espérance conditionnelle.

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{\infty} M_i \mathbb{1}_{\{i\leq N\}} \middle| N\right]\right]$$

Comme il n'est pas facile de manipuler une somme dont les bornes sont aléatoires, on la réécrit de la façon suivante.

$$= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{E}\left[M_i \mathbb{1}_{\{i\leq N\}} \middle| N\right]\right]$$

Puisque les termes de la somme sont positifs, les sommes partielles sont croissantes et on peut donc utiliser le théorème de convergence monotone.

$$= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{\{i\leq N\}} \mathbb{E}\left[M_i \middle| N\right]\right]$$

L'indicatrice étant N-mesurable, on peut la sortir de l'espérance.

$$= \mathbb{E}\left[\sum_{i=1}^{N} \mathbb{E}\left[M_i \middle| N\right]\right]$$

Maintenant que l'espérance est rentrée dans la somme, on peut réécrire la somme sous sa forme normale.

$$= \mathbb{E}\left[\sum_{i=1}^{N} m\right]$$

Comme les variables M sont indépendantes de N (ça c'est d'après la définition du modèle de Galton-Watson : le nombre d'enfants d'un individu ne dépend pas de son nombre de frères ou soeurs), ces espérances valent aussi m.

$$= \mathbb{E}[mN]$$

La somme n'est maintenant plus qu'un produit.

$$= m\mathbb{E}[N]$$

La constante m peut sortir de l'espérance.

$$= m \times m^k$$

$$= m^{k+1}$$

Il ne reste plus qu'à utiliser l'hypothèse de récurrence pour obtenir le résultat voulu !

Notez que toute l'astuce de ce calcul consiste à contourner habilement la variable N qui nous gêne dans les bornes de la somme. Ce calcul est en fait une démonstration dans un cas particulier d'un résultat qui s'appelle la formule de Wald du nom du mathématicien hongrois Abraham Wald, si vous connaissiez cette formule il suffisait de l'appliquer pour obtenir le résultat en une ligne.

Bien, ceux qui voulaient sauter le calcul vous pouvez rouvrir les yeux, et reprendre votre lecture à partir d'ici !


Abraham Wald
(1902-1950)

Le nombre total de descendants

Maintenant que nous savons combien il y a de descendants à la $k^{\text{ème}}$ génération en moyenne, il ne devrait plus être très dur de trouver la moyenne du **nombre total** de descendants dans l'arbre. Il suffit de faire une somme. Autrement dit, si on note S le nombre total d'individus dans l'arbre, on a :

$$\mathbb{E}[S] = \sum_{k=0}^{\infty} m^k.$$

(Et entre parenthèses pour ceux qui veulent la démonstration rigoureuse, il ne s'agit que d'une simple application du théorème de convergence monotone (encore lui !) :

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{k=0}^{\infty} N_k\right] = \sum_{k=0}^{\infty} \mathbb{E}[N_k] = \sum_{k=0}^{\infty} m^k.$$

où N_k désigne le nombre d'individus à la $k^{\text{ème}}$ génération. Je ferme la parenthèse.)

Eh, mais on reconnait la somme des termes d'une suite géométrique ! On sait calculer cette somme :

$$\sum_{k=0}^{\infty} m^k = \frac{1}{1-m}.$$

Euh... Pas trop vite. Vous êtes sûrs que vous n'oubliez pas quelque chose ?

Ah oui ! Pour que la somme des termes d'une suite géométrique soit finie, il faut évidemment que m soit strictement plus petit que 1. Et si $m \geq 1$ alors la somme est infinie. Autrement dit on a :

$$\mathbb{E}[S] = \begin{cases} \frac{1}{1-m} & \text{si } m < 1; \\ +\infty & \text{si } m \geq 1. \end{cases}$$

On voit donc apparaître deux cas distincts. Quand $m < 1$ on dit que l'arbre est **sous-critique**, quand $m > 1$ on dit qu'il est **sur-critique**. Et dans le cas intermédiaire où $m = 1$ qui est un peu spécial comme nous allons le voir par la suite, on dit qu'on est dans le cas **critique**.

En bref, on a :

Cas sous-critique $m < 1$	Cas critique $m = 1$	Cas sur-critique $m > 1$
$\mathbb{E}[S] = \frac{1}{1-m}$	$\mathbb{E}[S] = +\infty$	$\mathbb{E}[S] = +\infty$

Il y a une conclusion que l'on peut tout de suite tirer de ce tableau, c'est que dans le cas sous-critique, c'est-à-dire si le nombre moyen d'enfants par individu est strictement inférieur à 1, la descendance de l'arbre va forcément s'éteindre. En effet, si la moyenne du nombre d'individus dans l'arbre est finie, cela signifie que le nombre d'individus dans l'arbre ne peut jamais être infini. Or dire que l'arbre n'est pas infini c'est précisément dire que la descendance va s'éteindre.

Mais alors dans les cas critique et sur-critique on peut en conclure que l'arbre est infini et donc que la descendance ne s'éteint pas ? Non ?

Non. Là il y a une subtilité.

Quand une moyenne (ou une espérance) de nombres positifs est finie alors tous ses termes sont finis, **MAIS** quand la moyenne (ou espérance) est infinie **ça ne veut pas forcément dire qu'il y a un terme infini dans la moyenne !**

Un des exemples les plus classiques pour se convaincre de ceci est le jeu suivant. Au début du jeu on vous donne une pièce de 2€. Vous lancez cette pièce et tant que vous tombez sur "pile" votre somme est multipliée par 2 et le jeu continue. Le jeu s'arrête dès que vous tombez sur "face". Quelle est alors l'espérance de votre gain ?

- Avec probabilité 1/2, vous allez tomber sur face dès le premier lancer et votre gain n'est alors que de 2€ (la pièce initiale.)

- Avec probabilité 1/4, vous allez gagner 4€ en faisant d'abord un pile puis une face.
- Avec probabilité 1/8, vous allez gagner 8€ en faisant d'abord deux piles puis une face.
- ...
- Avec probabilité $1/2^n$, vous allez gagner 2^n€ en faisant d'abord $n-1$ piles puis une face.
- ...

Votre espérance de gain est alors égale à :

$$\mathbb{E}\left[\text{gain}\right] = \sum_{n=1}^{\infty} \frac{1}{2^n} \times 2^n = \sum_{n=1}^{\infty} 1 = +\infty.$$

Vous voyez que votre espérance de gain est infinie alors qu'il n'est pas possible de gagner une infinité d'euros puisque vous allez nécessairement faire une face à un moment donné.

Un autre exemple qui ne fait pas intervenir les probabilités est de se demander combien vaut la moyenne de tous les entiers naturels (0, 1, 2, 3, 4,...). Il est facile de se convaincre que leur moyenne est infinie (calculez la moyenne des n premiers entiers puis faites tendre n vers l'infini si vous n'êtes pas convaincu...) et pourtant tous les nombres entiers sont finis !

Bref, revenons à nos arbres. Tout ce que nous savons pour l'instant c'est que dans le cas sous-critique la descendance de l'ancêtre s'éteint et dans les deux autres cas, nous ne pouvons rien dire pour l'instant.

La fonction génératrice

Nous en arrivons maintenant à l'outil fondamental de ce cours : **la fonction génératrice**. Il s'agit une fonction qui va nous permettre d'explorer l'arbre en faisant passer certaines propriétés d'une génération à la suivante. Il est donc très important que vous compreniez sa signification pour pouvoir aborder le raisonnement qui viendra ensuite. Bref, c'est le moment d'être attentif !

Définition

Commençons par une définition heuristique avant de voir comment cela se traduit en formules mathématiques.

Supposons que l'on veuille étudier une certaine propriété de l'arbre. Par exemple, cette propriété peut être quelque chose du genre :

- « L'ancêtre a trois enfants » ;
- « Personne n'a plus de 3 enfants dans les 10 premières générations » ;
- « La descendance de l'ancêtre va s'éteindre ».

On peut aussi imaginer qu'on rajoute au modèle le fait que chaque individu a un sexe, féminin ou masculin, et la propriété peut alors être :

- « L'ancêtre est une fille » ;
- « Il y a au moins dix garçons dans les cinq premières générations » ;
- « Tous les petits-enfants de l'ancêtre sont des filles ».

Bref, on peut imaginer toute sorte de propriétés. Chacune d'entre elles possède une certaine probabilité.

La fonction génératrice g est alors la fonction qui a la probabilité s d'une propriété associe la probabilité que **tous les enfants de l'ancêtre vérifient cette propriété.**

$$\begin{array}{rccl} g: & [0,1] & \to & [0,1] \\ & s & \mapsto & \mathbb{P}\left[\text{tous les enfants de l'ancêtre vérifient une propriété de probabilité } s\right] \end{array}$$

Mais non ne fuyez pas, vous allez voir, avec les exemples ça passe mieux. Reprenons donc nos exemples de propriétés.

Si s est la probabilité pour que...	...alors $g(s)$ est la probabilité pour que...
l'ancêtre ait trois enfants	tous les enfants de l'ancêtre aient trois enfants.
personne n'ait plus de 3 enfants dans les 10 premières générations	dans les descendances des enfants de l'ancêtre, personne n'ait plus de 3 enfants dans les 10 premières générations. *On peut reformuler cette propriété de la façon suivante : de la génération 1 à la génération 11, personne n'a plus de 3 enfants.*
la descendance de l'ancêtre s'éteigne	les descendances de tous les enfants de l'ancêtre s'éteignent.
l'ancêtre soit une fille	tous les enfants de l'ancêtre soient des filles.
il y ait au moins dix garçons dans les cinq premières générations	tous les enfants de l'ancêtre aient au moins dix garçons dans les cinq générations après eux.
tous les petits-enfants de l'ancêtre soient des filles	tous les arrière-petits-enfants de l'ancêtre soient des filles.

C'est bien beau tout ça mais comment on la calcule concrètement la fonction génératrice ?

Soit P une propriété de probabilité s nous cherchons donc à calculer la probabilité que tous les enfants de l'ancêtre vérifient cette propriété. On a donc

$$g(s) = \mathbb{P}\left[\text{tous les enfants de l'ancêtre vérifient } P\right] = \sum_{n=0}^{\infty} \mathbb{P}\left[\text{l'ancêtre a } n \text{ enfants qui vérifient } P\right]$$

On sait que la probabilité pour que l'ancêtre ait n enfants est égale à p_n, et la probabilité que chacun d'entre eux vérifie P est alors égale à s^n. Au final, on a donc :

$$g(s) = \sum_{n=0}^{\infty} p_n s^n$$

Voilà notre belle fonction génératrice qui va nous être bien utile ! Maintenant que nous avons sa formule, étudions la un peu...

L'étude de la fonction génératrice

Le domaine de définition de g est bien entendu $[0,1]$ puisque l'argument de g est une probabilité. On peut donc commencer par calculer les valeurs de g aux bornes de cet intervalle. On a :

$$g(0) = \sum_{n=0}^{\infty} p_n 0^n = p_0$$

En effet, tous les termes sont nuls sauf le premier car $s^0 = 1$. Calculons maintenant $g(1)$:

$$g(1) = \sum_{n=0}^{\infty} p_n 1^n = \sum_{n=0}^{\infty} p_n = 1$$

Une remarque pour les étudiants post-bac. Notez que le fait que $g(1)$ soit bien défini, c'est-à-dire que la somme converge pour $s = 1$ prouve que le rayon de convergence de la série entière est supérieur à 1 et donc que g est bien définie sur tout notre intervalle et qu'elle y est infiniment dérivable. Nous avons donc bien le droit de faire toutes les opérations que nous allons faire maintenant.

Pour connaitre le sens de variation de g on calcule sa dérivée :

$$g'(s) = \sum_{n=0}^{\infty} p_n n s^{n-1} = p_1 + 2p_2 s + 3p_3 s^2 + 4p_4 s^3 + 5p_5 s^4 + \dots$$

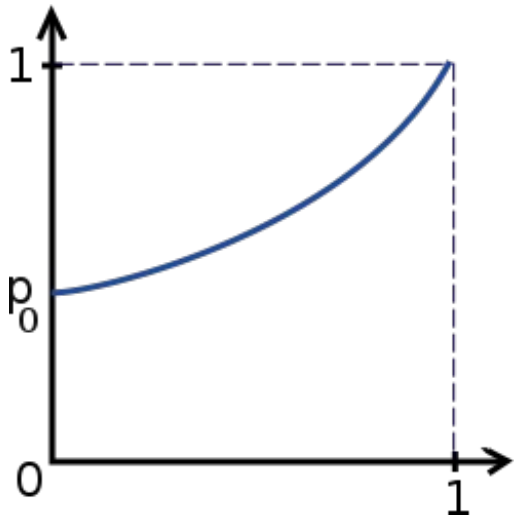
On remarque alors que tous les termes de cette somme sont positifs, la dérivée est donc positive et g est croissante.

Poussons l'étude un peu plus loin en regardant la dérivée seconde :

$$g''(s) = \sum_{n=0}^{\infty} p_n n(n-1) s^{n-2} = 2p_2 + 6p_3 s + 12p_4 s^2 + 20p_5 s^3 + \dots$$

Encore une fois, tous les termes sont positifs donc la dérivée seconde est positive ce qui veut dire que g est convexe. (Si vous ne connaissez pas ce mot, dire qu'une fonction est convexe signifie simplement que la courbe du graphe est « arrondie vers le bas ». Vous trouverez plus d'informations à ce sujet sur la page Wikipedia sur les fonctions convexes.)

En résumé, le graphe de g ressemble à ceci :



Notez que g prend bien ses valeurs dans $[0,1]$ ce qui est normal puisque $g(s)$ désigne aussi une probabilité !

Disparaitre ou survivre ?

Il est temps de revenir à notre problème initial ! Nous sommes maintenant fin prêts pour répondre à la question de Galton.

Tout le raisonnement tient dans une seule phrase : *dire que la descendance de l'ancêtre s'éteint, c'est exactement la même chose que de dire que les descendances de tous les enfants de l'ancêtre s'éteignent !*

Appelons q la probabilité d'extinction, c'est-à-dire la probabilité que l'ancêtre n'ait qu'un nombre fini de descendants. Alors la probabilité pour que les descendances de tous les enfants de l'ancêtre s'éteignent est égale à $g(q)$. Mais comme les deux événements sont identiques, on a donc :

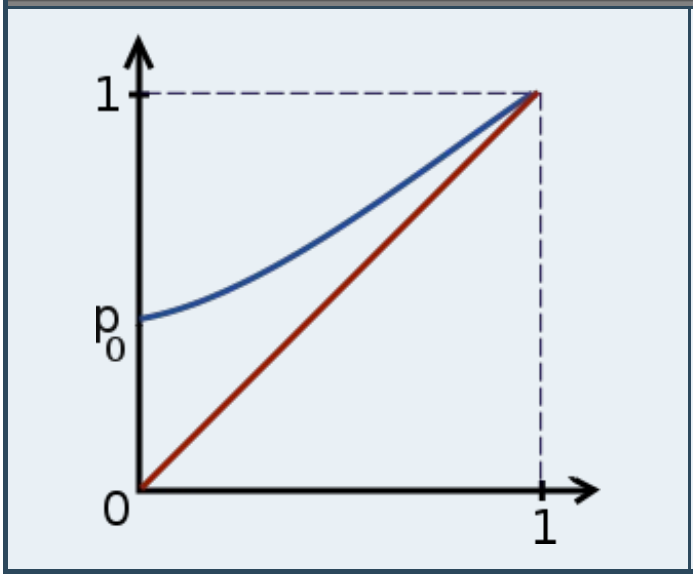
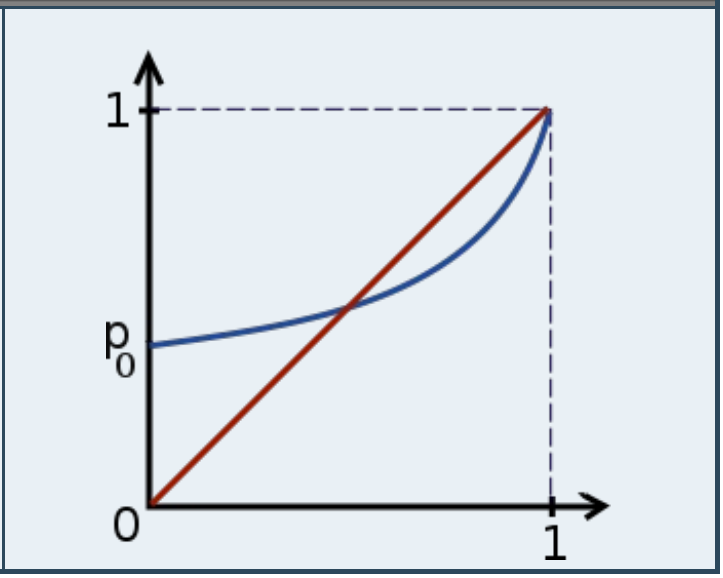
$$g(q) = q.$$

La probabilité q est donc solution de l'équation $g(s) = s$! Pour trouver q il nous suffit donc de résoudre cette équation. Astucieux, n'est-ce pas ?

Graphiquement, trouver une solution à l'équation $g(s) = s$, cela revient à trouver un point d'intersection entre le graphe de g et

la droite d'équation $f(s) = s$.

Comme nous savons que g est croissante et convexe, il y a deux cas de figure possibles :

Soit le graphe de g est au-dessus de la droite $f(s) = s$ et il n'y a qu'un seul point d'intersection en $s = 1$.	Soit le graphe de g traverse la droite et il y a alors deux solutions : s=1 et une autre strictement plus petite que 1.
	

Mais comment savoir laquelle de ces deux situations est la bonne ?

Pour répondre à cette question il va falloir faire preuve d'un peu d'astuce. En fait, il suffit de regarder ce qui se passe en $s = 1$:

- si le graphe de g arrive en $s = 1$ en étant *au-dessus* de la diagonale, alors nous sommes dans le premier cas ;
- si au contraire le graphe de g arrive en $s = 1$ en étant *en dessous* de la diagonale, alors nous sommes dans le deuxième cas.

Et pour savoir ça, il suffit de calculer la dérivée de g en $s = 1$. Si $g'(1) < 1$, c'est-à-dire si la tangente à la courbe en 1 a un coefficient directeur plus petit que 1, nous sommes dans le premier cas. Si au contraire $g'(1) > 1$, alors la pente en 1 est plus grande que 1 et nous sommes dans le deuxième cas.

Nous avons déjà calculé la dérivée de g, elle vaut :

$$g'(s) = \sum_{n=0}^{\infty} n p_n s^{n-1}.$$

Et donc,

$$g'(1) = \sum_{n=0}^{\infty} n p_n.$$

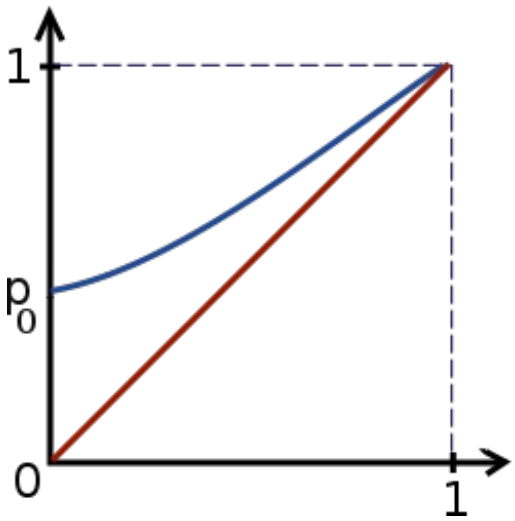
Vous le reconnaissez ce nombre-là ? Mais oui, c'est m ! Le nombre moyen d'enfants. Comme par hasard le revoilà qui pointe le bout de son nez : $g'(1) = m$!

Nous en revenons donc à la même conclusion que précédemment : les arbres ne grandissent pas de la même façon selon que m est plus grand ou plus petit que 1. Nous pouvons donc maintenant étudier les trois cas que nous avions déjà mis en avant

(sous-critique, critique et sur-critique) plus en détail.

Le cas sous-critique

Commençons par le cas sous-critique, c'est-à-dire $m < 1$. Dans ce cas, comme nous l'avons dit, la fonction génératrice reste au-dessus de la droite $f(s) = s$:



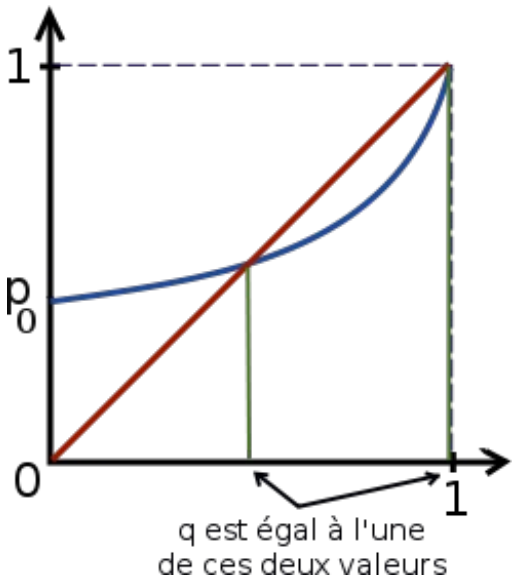
Il n'y a qu'un seul point d'intersection avec la droite $f(s) = s$, qui se trouve en $s = 1$. La probabilité d'extinction est donc égale à 1.

Remarquez que nous n'apprenons rien puisque le cas sous-critique est le seul dans lequel nous avions déjà trouvé la réponse grâce aux calculs d'espérance ci-dessus.

$$q = 1$$

Le cas sur-critique

Dans le cas sur-critique $m > 1$, la fonction génératrice et la droite $f(s) = s$ ont un autre point d'intersection d'abscisse strictement plus petite que 1. Nous avons donc de l'espoir pour que la descendance de l'ancêtre ne s'éteigne pas ! 



Ceci dit pour l'instant, nous savons simplement que la probabilité d'extinction q correspond à l'un des deux points d'intersection,

mais rien ne nous prouve que ce ne soit pas quand même 1. Rassurez vous nous allons bien montrer que q correspond à la solution plus petite que 1 de $f(s) = s$, mais il faut pour cela faire un raisonnement un peu plus précis.

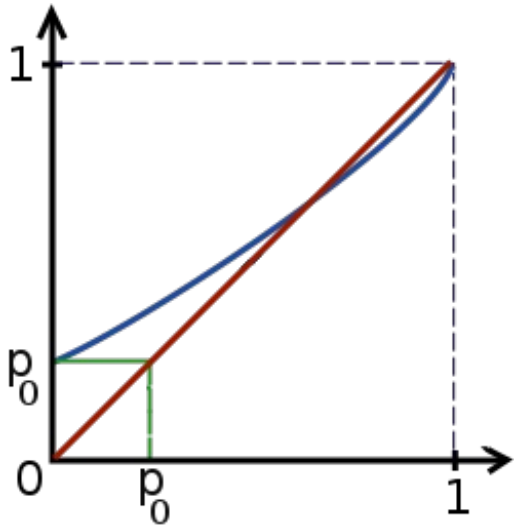
Nous allons calculer pour chaque génération la probabilité pour que la descendance de l'ancêtre soit déjà éteinte à cette génération.

Génération 1. Quelle est la probabilité pour que la descendance de l'ancêtre s'éteigne dès la génération 1 ? C'est tout simplement la probabilité pour qu'il n'ait pas d'enfants, c'est-à-dire p_0.

Génération 2. Quelle est la probabilité pour que la descendance de l'ancêtre soit éteinte à la génération 2, c'est-à-dire la probabilité qu'il n'ait pas de petits-enfants ? C'est la probabilité qu'aucun de ses enfants n'ait d'enfants, autrement dit cette probabilité et égale à $g(p_0)$.

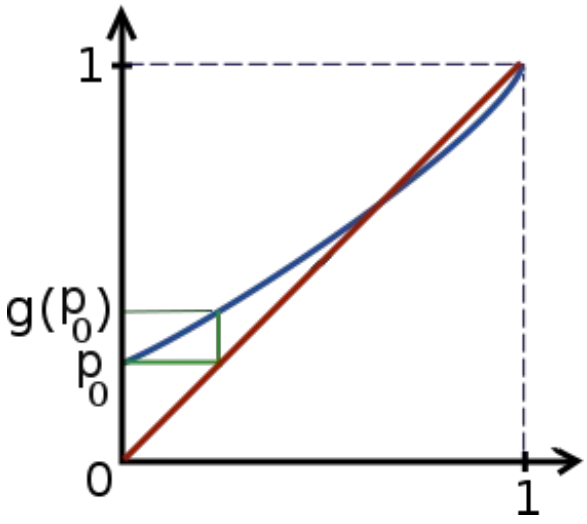
Pour placer $g(p_0)$ sur le graphique nous allons utiliser une construction astucieuse et très classique quand on veut itérer une fonction. Remarquez que p_0 est déjà porté sur le graphique, c'est l'ordonnée à l'origine du graphe de g, cependant pour trouver $g(p_0)$, il faudrait que p_0 soit porté sur l'axe des abscisses et non sur celui des ordonnées.

Pour cela, nous allons projeter horizontalement le point d'ordonnée p_0 sur la droite d'équation $f(s) = s$. De cette façon, on obtient un point dont l'abscisse est égale à l'ordonnée, c'est-à-dire à p_0.



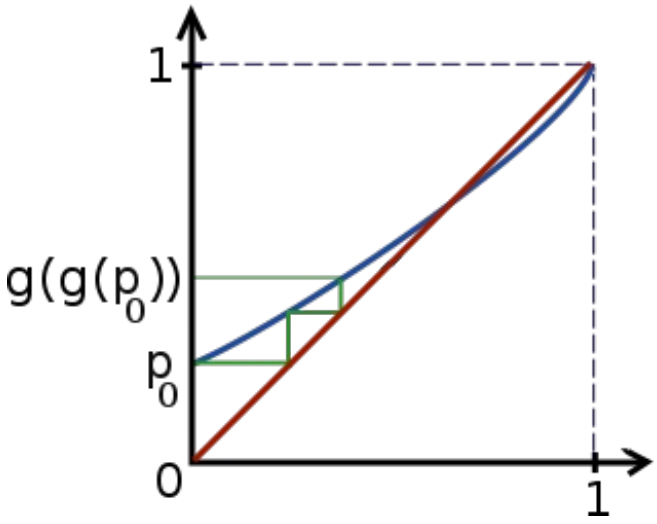
(Notez que je n'ai pas pris exactement la même fonction g que dans les exemples précédents de façon à ce que la construction qui va suivre soit plus visible sur le graphique. Mais du moment que nous sommes toujours dans le cas sur-critique, ça ne change rien au raisonnement.)

Il ne nous reste maintenant plus qu'à repartir verticalement sur le graphe de g pour obtenir la valeur de $g(p_0)$:



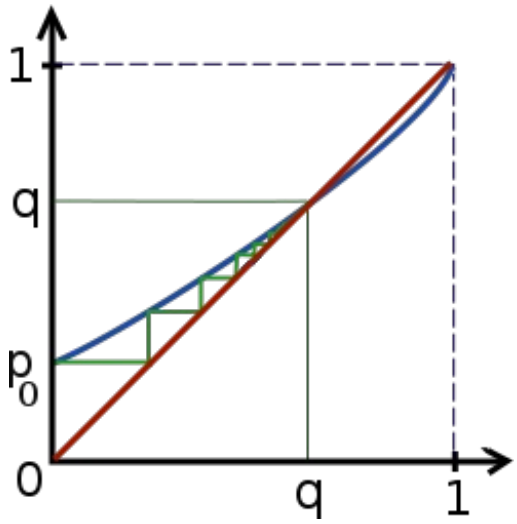
Génération 3. Quelle est la probabilité pour que la descendance de l'ancêtre soit éteinte à la génération 3 ? C'est la probabilité pour que la descendance de chacun de ses enfants soit éteinte à la génération 2. Autrement dit, c'est $g(g(p_0))$.

Sur le graphe, il suffit donc de reproduire ce que nous avons fait à la génération 2, mais en partant cette fois de $g(p_0)$:



Générations suivantes... Vous commencez à comprendre le principe. La probabilité pour que la descendance de l'ancêtre soit éteinte à la génération k est égale à la probabilité que les descendances de tous ses enfants soient éteintes après $k-1$ générations. Par récurrence cette probabilité vaut donc $g^{k-1}(p_0)$.

Sur le graphique on construit la suite $g^{k-1}(p_0)$ en utilisant toujours le même procédé :



Et là c'est le soulagement : cette suite tend bien vers la solution strictement inférieure à 1 de $g(s) = s$! Autrement dit $q < 1$ ce qui signifie encore que la probabilité de survie de la descendance de l'ancêtre est strictement positive.

Bon, les preuves sur les dessins c'est bien joli, mais pour être plus rigoureux nous allons également prouver ce résultat par le calcul. Encore une fois, vous pouvez sauter si vous le souhaitez.

Pour simplifier notons x la solution strictement inférieure à 1 de $g(s) = s$. En bref, nous voulons montrer que $q = x$.

Nous allons prouver par récurrence que tous les $g^k(p_0)$ sont plus petits que x.

C'est vrai pour $k = 0$. En effet on a $0 \leq x$ donc comme g est croissante, $g(0) \leq g(x)$ c'est-à-dire $p_0 \leq x$. Ainsi on a bien $g^0(p_0) = p_0 \leq x$. La récurrence est initialisée.

Supposons maintenant que $g^k(p_0) \leq x$, alors de la croissance de g on déduit que $g(g^k(p_0)) \leq g(x)$, c'est-à-dire $g^{k+1}(p_0) \leq x$ puisque $x = g(x)$. Ce qui achève la récurrence.

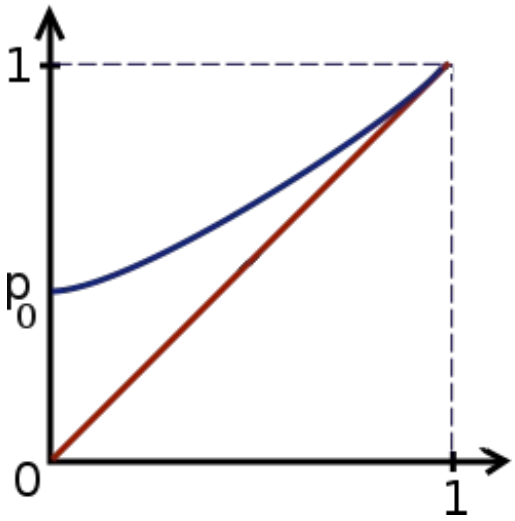
Ainsi les $g^k(p_0)$ sont tous inférieurs à x et leur limite q est donc elle aussi inférieure ou égale à x. Ainsi q ne peut pas être égale à 1 ! On a donc forcément $q = \lim_{k\to\infty} g^k(p_0) = x < 1$. Et le tour est joué.

$$q < 1$$

Le cas critique

Reste le cas critique. Que se passe-t-il si $m = 1$?

Dans ce cas, la droite $f(s) = s$ est la tangente en 1 au graphe de g.



Comme la fonction génératrice g est convexe, on voit que comme dans le cas sous-critique, la seule solution à l'équation $g(s) = s$ est $s = 1$. Par conséquent, la probabilité d'extinction q est égale à 1 !

Remarquez que le cas critique justifie la méfiance que nous avions quand nous avions calculé l'espérance du nombre de descendants de l'ancêtre. Vous voyez que dans ce cas là, l'espérance du nombre de descendants est infinie et pourtant la descendance s'éteint de façon certaine.

En réalité, il y a tout de même un cas particulier. Lors de l'étude de g nous avons vu que cette fonction est convexe, mais elle n'est pas forcément *strictement* convexe. Il est tout à fait possible que $g'' = 0$ si tous les p_n valent 0 à partir de $n = 2$. Dans ce cas, le graphe de g est une droite $g(x) = p_0 + p_1 x$. Et comme de plus, la dérivée en 1 est égale à 1, on a en réalité $p_0 = 0$ et $p_1 = 1$ autrement dit, $g(s) = s$.

Dans ce cas le graphe de g se confond avec la droite et tous les nombres sont solutions de l'équation $g(s) = s$!

Heureusement ce cas est en réalité très simple à étudier. En effet, si $p_1 = 1$ et que tous les autres p_n valent 0 cela signifie que tous les individus n'ont qu'un seul enfant avec probabilité 1. Autrement dit à chaque génération il n'y a qu'un seul individu et la descendance ne s'éteint pas donc $q = 0$.

Évidemment ce cas dégénéré ne correspond pas à des situations concrètes, mais il était nécessaire de le mentionner quand même pour que l'étude mathématique du modèle soit complète.

En bref, dans le cas critique on a :

si $p_1 < 1$	si $p_1 = 1$
$q = 1$	$q = 0$

Récapitulons

En bref, nous pouvons résumer les résultats que nous avons obtenus dans les trois cas par le tableau suivant :

	Cas sous-critique	Cas critique	Cas sur-critique
Nombre moyen d'enfants	$m < 1$	$m = 1$	$m > 1$
Nombre moyen de descendants	$\mathbb{E}[S] = \frac{1}{1-m}$	$\mathbb{E}[S] = +\infty$	$\mathbb{E}[S] = +\infty$
Probabilité d'extinction	$q = 1$	$q = \begin{cases} 1 & \text{si } p_1 < 1 \\ 0 & \text{si } p_1 = 1 \end{cases}$	$q < 1$

Application des résultats

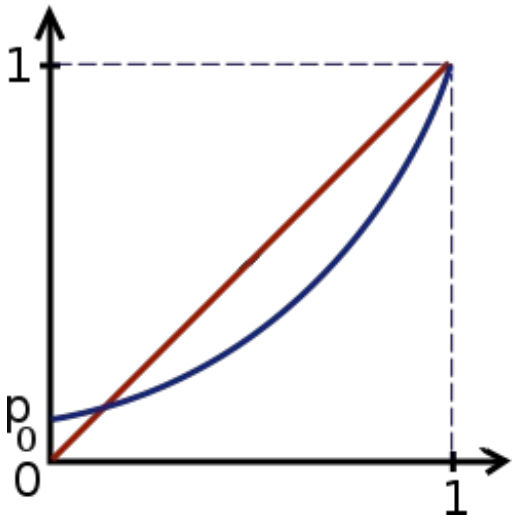
C'est bon, vos neurones ne chauffent pas trop ? Dans cette partie ils vont pouvoir se reposer un petit peu (pas trop quand même) puisque nous allons juste mettre en application les résultats que nous avons obtenus dans l'exemple concret donné par les statistiques des années 60.

Probabilité d'extinction de la descendance

La fonction génératrice dans ce cas particulier est égale à :

$$g(s) = 0,10 + 0,18s + 0,40s^2 + 0,22s^3 + 0,07s^4 + 0,03s^5.$$

On a déjà calculé que dans ce cas là $m = 2,07$, nous sommes donc dans le cas sur-critique, ce qui se confirme en traçant le graphe de g :



Il y a bien une solution strictement inférieure à 1 à l'équation $g(s) = s$ et on voit même qu'elle est assez petite ce qui est assez logique puisque le nombre moyen d'enfants est plus de deux fois supérieur au seuil critique de 1.

Pour trouver une valeur approchée de q, suffit de se rappeler que

$$q = \lim_{k \to +\infty} g^k(p_0)$$

En prenant k suffisamment grand, on obtient une bonne approximation de q :

$$q \approx 0,130944...$$

Autrement dit la probabilité pour que la descendance d'un individu s'éteigne est égale à environ 13%.

Notez que dans ce cas il est possible de calculer une valeur exacte de q, mais à vrai dire ce n'est pas vraiment une bonne idée car les calculs sont assez ignobles. Pour tout vous dire, la valeur exacte de q est

$$q = -\frac{5}{6} - \frac{1}{6}\sqrt{-39 - 748\sqrt[3]{\frac{2}{40025+3\sqrt{201250569}}} + 2^{2/3}\sqrt[3]{40025+3\sqrt{201250569}}} + \frac{1}{2}\sqrt{-\frac{26}{3} + \frac{748}{9}\sqrt[3]{\frac{2}{40025+3\sqrt{201250569}}} - \frac{1}{9}2^{2/3}\sqrt[3]{40025+3\sqrt{201250569}} + \frac{586}{9\sqrt{\sqrt{-39-748\sqrt[3]{\frac{2}{40025+3\sqrt{201250569}}}+2^{2/3}\sqrt[3]{40025+3\sqrt{201250569}}}}}}$$

Vérifiez si vous voulez !

Probabilité d'extinction des noms de famille

Le problème originel de Galton n'était pas tout à fait le même. En effet, il s'agissait de savoir si *les noms de familles* des lords anglais disparaissaient. Dans ce cas, il ne faut pas étudier l'ensemble de la descendance des individus, mais seulement leur descendance mâle. En effet, à cette époque seuls les hommes transmettaient leurs noms de famille tandis que les femmes en changeaient en se mariant.

Aujourd'hui, même si elle est encore majoritairement suivie, la tradition des noms de familles transmis par le père n'est plus obligatoire. Un enfant peut aussi bien recevoir le nom de son père ou de sa mère. Cependant, ça ne change rien au problème : pour un individu, seul un enfant sur deux en moyenne va recevoir son nom.

Dans la suite, pour simplifier les explications je vais faire comme si les noms ne se transmettaient que de père en fils, comme autrefois. C'est juste parce qu'il est plus cours d'écrire *"fils"* que *"enfants qui ont reçu le même nom de famille"* à chaque fois. Mais les calculs sont les mêmes dans les deux cas.

Si on reprend l'exemple des années 60, cela signifie que le nombre moyen d'enfants qui vont porter le même nom de famille qu'un individu donné n'est plus égal à 2,07 mais à la moitié : $m = 1,035$.

Nous sommes donc bien encore dans le cas sur-critique, mais c'est vraiment limite !

Reprenons donc notre exemple préféré des années 60. Quelle est la probabilité pour que le nom de famille d'un individu s'éteigne selon ces statistiques ?

Pour répondre à cette question, il nous faut faire quelques calculs préliminaires pour calculer les nouvelles valeurs des nouveaux p_n dans ce cas là. Pour ne pas nous embrouiller nous allons leur donner un nouveau nom : q_n. Autrement dit q_n est la probabilité pour un individu d'avoir n garçons.

On a alors

$$q_n = \sum_{i=0}^{\infty} p_i \times \binom{i}{n} \times \frac{1}{2^i}$$

Cette formule mérite quelques explications vous ne croyez pas ?

Pour chaque i on sait qu'un individu a une probabilité p_i d'avoir i enfants. Alors la probabilité d'avoir n fils parmi ces i enfants est égale à $\binom{i}{n} \times \frac{1}{2^i}$. La combinaison $\binom{i}{n} = \frac{i!}{n!(i-n)!}$ correspond au nombre de façons de choisir les n fils parmi les i enfants, et $\frac{1}{2^i} = \left(\frac{1}{2}\right)^n \times \left(\frac{1}{2}\right)^{i-n}$ est la probabilité pour que les n choisis soient effectivement des garçons et que les $i-n$ autres soient des filles.

Notez que si $n > i$, alors par convention $\binom{n}{i}$ est égal à 0. Il n'est pas possible d'avoir plus de fils que d'enfants !

Si vous avez encore du mal à comprendre cette formule, prenons un exemple. Calculons q_2, dans le cas des années 60. Il y a plusieurs façons d'avoir deux fils, voyons les différents cas :

- On peut avoir deux enfants qui sont tous les deux des fils. Ceci se produit avec probabilité $p_2/4$ car il y a une probabilité p_2 d'avoir deux enfants et $(1/2) \times (1/2)$ que ce soient deux garçons.
- On peut avoir trois enfants dont deux fils et une fille. Ceci se produit avec probabilité $p_3 \times 3/8$ car il y a une probabilité p_3 d'avoir 3 enfants, il y a trois choix possibles pour les deux garçons parmi ces trois enfants (ils peuvent être le premier et le deuxième ou bien le premier et le troisième ou bien le deuxième et le troisième) puis une probabilité $(1/2) \times (1/2) \times (1/2)$ que chacun de ces scénarios se produisent.
- On peut avoir quatre enfants dont deux fils et deux filles. Ceci se produit avec probabilité $p_4 \times 6/16$ car il y a une probabilité p_4 d'avoir 4 enfants, il y a six choix possibles pour les deux garçons parmi ces quatre enfants puis une probabilité $(1/2) \times (1/2) \times (1/2) \times (1/2)$ que chacun de ces scénarios se produisent.
- On peut avoir cinq enfants, dont deux fils et trois filles. Ceci se produit avec probabilité $p_5 \times 10/32$ car il y a une probabilité p_5 d'avoir 5 enfants, il y a dix choix possibles pour les deux garçons parmi ces cinq enfants puis une probabilité $(1/2) \times (1/2) \times (1/2) \times (1/2) \times (1/2)$ que chacun de ces scénarios se produisent.

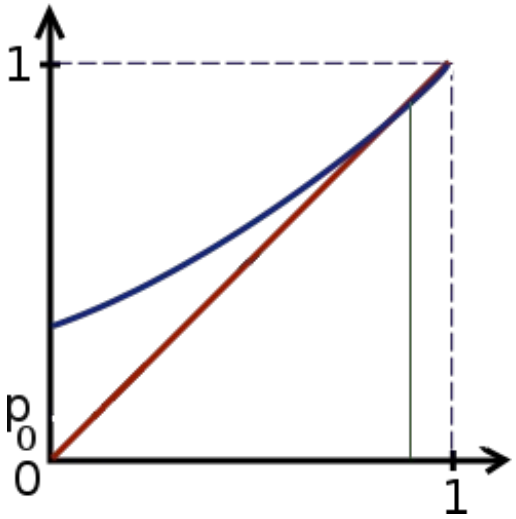
Et on s'arrête là car il n'est pas possible d'avoir plus de cinq enfants. Finalement, on trouve :

$$q_2 = \frac{p_2}{4} + \frac{3p_3}{8} + \frac{6p_4}{16} + \frac{10p_5}{32} = 0,218125.$$

En appliquant la même formule on trouve tous les q_n (essayez de les calculer si vous avez un doute) :

- $q_0 = 0,3228125$
- $q_1 = 0,3946875$
- $q_2 = 0,218125$
- $q_3 = 0,054375$
- $q_4 = 0,0090625$
- $q_5 = 0,0009375$

Vous pouvez remarquer que la somme de ces nombres est bien égale à 1 ! Si on trace la nouvelle fonction génératrice, on obtient ceci :



Argh. On voit qu'on est vraiment très près du cas critique ! Cette fois, on obtient :

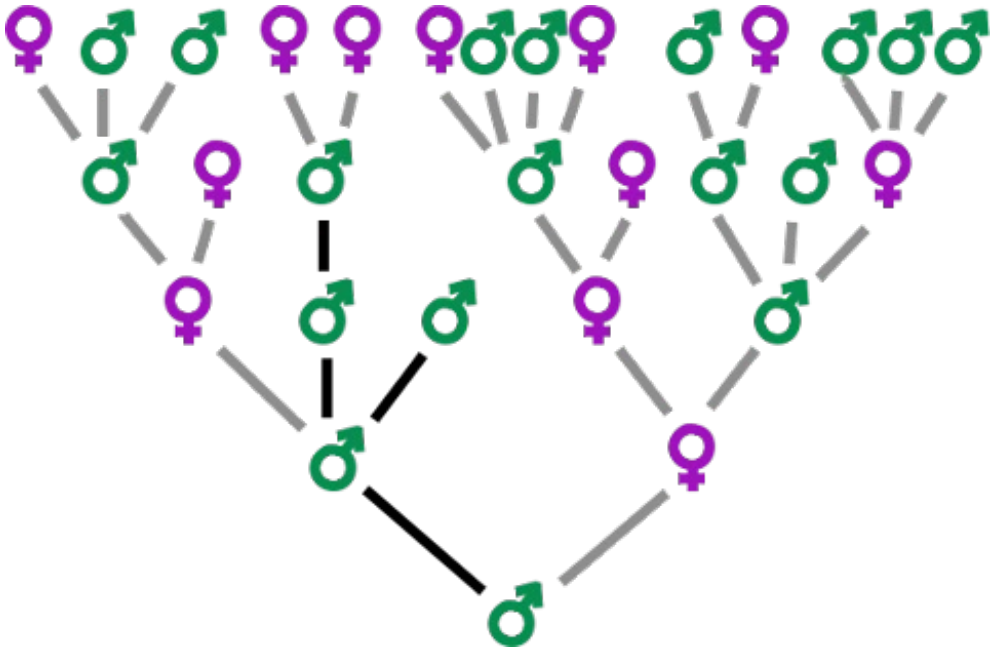
$$q \approx 0,919923...$$

La probabilité d'extinction d'un nom de famille à partir d'un de ses représentants est donc environ égale à 92% !

Conclusion

Finalement, on arrive à la conclusion étonnante dans notre exemple concret que la probabilité pour que la descendance d'un individu survive est assez élevée (86%) mais que la probabilité pour que son nom de famille survive est elle assez faible (8%).

Ainsi, une généalogie typique ressemble un peu à ça :



On voit que le nombre de descendants grandit rapidement en quelques générations de sorte que la descendance totale va survivre tandis que si on ne regarde que la descendance mâle directe, elle stagne à un ou deux individus pour finalement s'éteindre à la quatrième génération.
Ce cours sur les arbres de Galton-Watson s'arrête ici. N'hésitez pas à me laisser vos commentaires si vous avez la moindre remarque, incompréhension ou suggestion à propos de celui-ci.