

La simulation probabiliste avec Excel

Emmanuel Grenier emmanuel.grenier@isab.fr

I. Introduction

Incontournable lorsqu'il s'agit de gérer des phénomènes aléatoires complexes, la simulation probabiliste s'impose également dans l'enseignement des probabilités et de la statistique décisionnelle parce qu'elle permet d'aborder ces disciplines réputées théoriques et ardues par la voie de l'expérimentation. Voir le manuel du groupe « Le Cercle d'Excel'Ense » [1].

Après un rappel sur la manière de produire des valeurs pseudo-aléatoires avec Excel, nous montrerons qu'il est possible de simuler un large éventail de lois à partir de ces valeurs en utilisant une méthode simple et générale : la méthode des fractiles.

Nous présenterons ensuite des méthodes plus particulières, qui constitueront autant d'expériences à usage pédagogique.

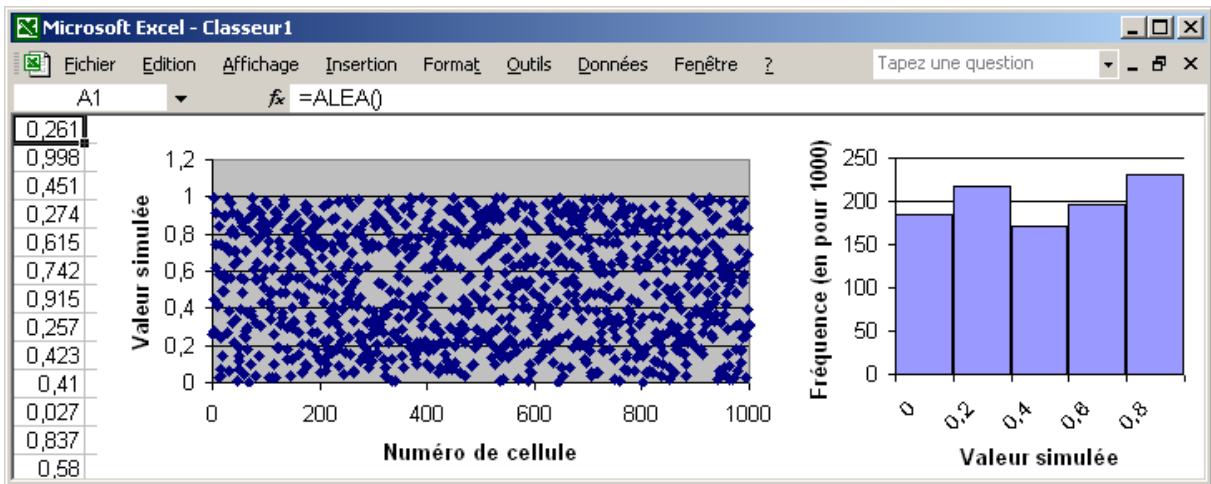
Le manuel du groupe « Le Cercle d'Excel'Ense » présente des applications dans la théorie des probabilités et en statistique décisionnelle : vérification de propriétés, caractérisation de la distribution de statistiques d'échantillonnage et étude des propriétés d'un estimateur.

Pour compléter, nous traiterons ici un cas pratique, pris dans le domaine de l'analyse du risque.

II. La production de valeurs pseudo-aléatoires avec Excel

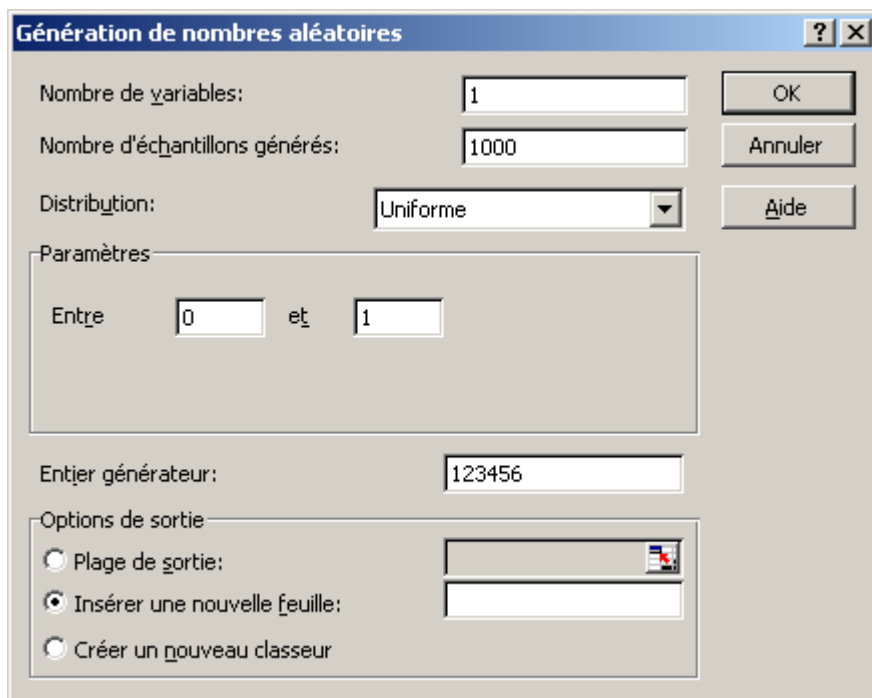
Imaginons un dispositif où une corde est tendue jusqu'à ce qu'elle casse. La rupture peut avoir lieu sur n'importe quelle portion de la corde, aussi bien sur les 20 premiers centimètres que sur les 20 derniers (pour simplifier, nous dirons que la corde fait une unité de longueur, soit 1 m de long). Sans information sur l'état de la corde, nous affectons la même probabilité à n'importe quelle portion de même longueur. Par exemple, la probabilité est la même sur la tranche de 0 à 0,2 m, c'est-à-dire sur les 20 premiers centimètres, que sur celle de 0,2 à 0,4 m, etc. C'est la loi « uniforme », que nous pouvons simuler avec Excel à partir de la fonction **ALEA** (voir dans le manuel la fiche « Fonction » correspondante et le chap. « Probabilités et jugement sur échantillon »)

Pour le vérifier, recopions la formule **=ALEA()** sur 1000 cellules et représentons la distribution des valeurs obtenues par un nuage de points ou par un histogramme (voir dans le manuel la fiche « Comment faire »).



On a à peu près autant de valeurs simulées entre 0 et 0,2 qu'entre 0,2 et 0,4, etc.

Nous aurions pu produire une série équivalente en passant par l'utilitaire **Génération de nombres aléatoires** (Allez dans le menu **Outils**. Si l'**Utilitaire d'analyse** n'apparaît pas, installez le en passant par **Macros complémentaires**)



Par rapport à la fonction **ALEA**, dont le résultat est volatile, l'utilitaire permet de reproduire la même série. Il suffit d'entrer le même **Entier générateur** (ici **123456**).

III. Simulation d'une loi par la méthode des fractiles

A. Domaine d'application

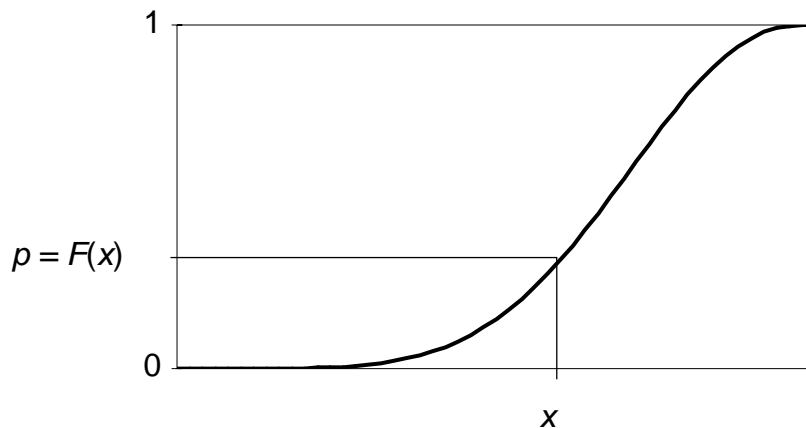
Toute loi s'il est possible de calculer la réciproque de sa fonction de répartition.

B. Principe

La fonction de répartition F d'une variable X correspond aux probabilités cumulées (voir le chap. « Probabilités et jugement sur échantillon ») :

$$F(x) = P(X \leq x).$$

Considérons une valeur possible x de la variable X et p la valeur correspondante de la fonction de répartition.



Si x est une réalisation quelconque de la variable X , on n'a pas d'information sur p , mis à part que sa valeur est comprise entre 0 et 1. On peut par conséquent considérer que p est la réalisation d'une variable de loi uniforme entre 0 et 1.

Réciproquement, si p est la réalisation d'une loi uniforme entre 0 et 1, le « fractile » correspondant, $x = F^{-1}(p)$, peut être considéré comme une réalisation de la variable X (pour une démonstration plus formelle voir par exemple l'ouvrage de G. Saporta [2]).

Avec Excel, on peut simuler la réalisation d'une variable de loi uniforme en utilisant la fonction **ALEA**. Par conséquent, on simule une réalisation de la variable X en appliquant la réciproque de sa fonction de répartition au résultat de la fonction **ALEA**.

C. Vérification sur un exemple

Prenons la loi normale standard (voir le chap « Probabilités et jugement sur échantillon »).

La fonction de répartition est calculée par la fonction **LOI.NORMALE.STANDARD**, sa réciproque par la fonction **LOI.NORMALE.STANDARD.INVERSE**. (voir les fiches « Fonction »)

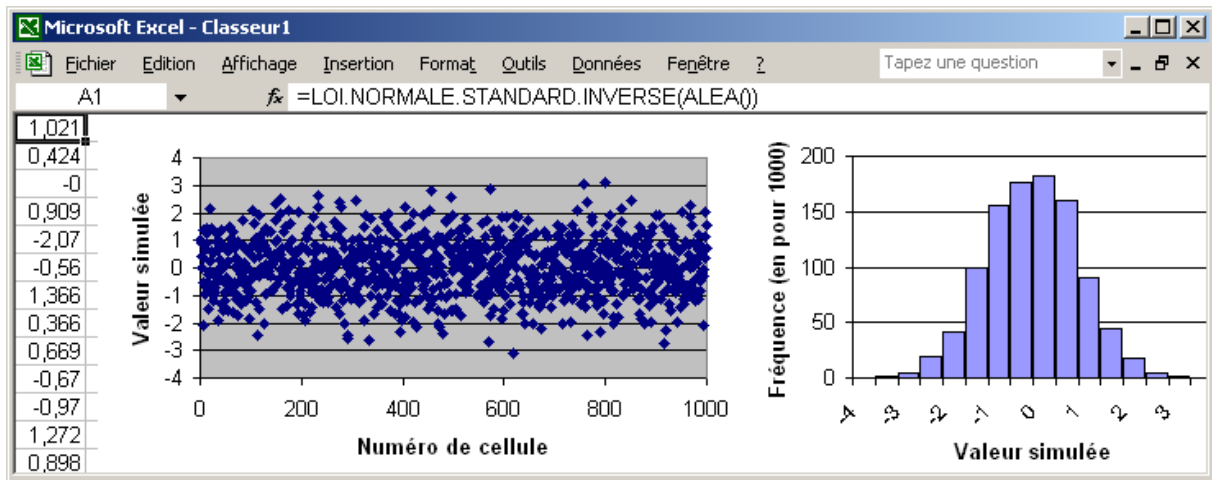
Par exemple, la formule **=LOI.NORMALE.STANDARD(1,96)** donne **0,975**, probabilité d'obtenir une valeur inférieure à 1,96.

En appliquant la fonction réciproque sur le résultat, c'est-à-dire en tapant la formule **=LOI.NORMALE.STANDARD.INVERSE(0,975)**, on retrouve la valeur de départ, **1,96**.

Appliquons la réciproque de la fonction de répartition au résultat de la fonction **ALEA**.

$$\mathbf{=LOI.NORMALE.STANDARD.INVERSE(ALEA())}$$

Recopions la formule sur 1000 cellules et représentons la distribution des valeurs obtenues par un nuage de points ou par un histogramme.



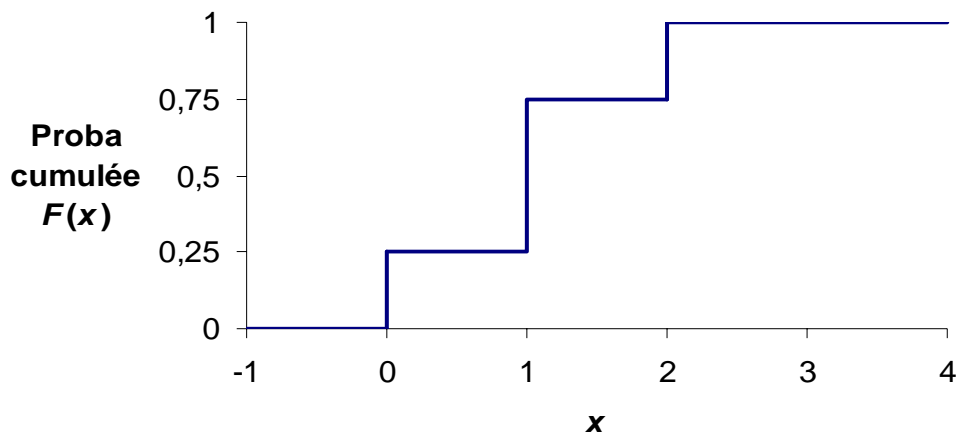
La distribution des valeurs simulées correspond bien à la loi normale standard.

D. Le cas particulier des variables discontinues

1. Méthode générale

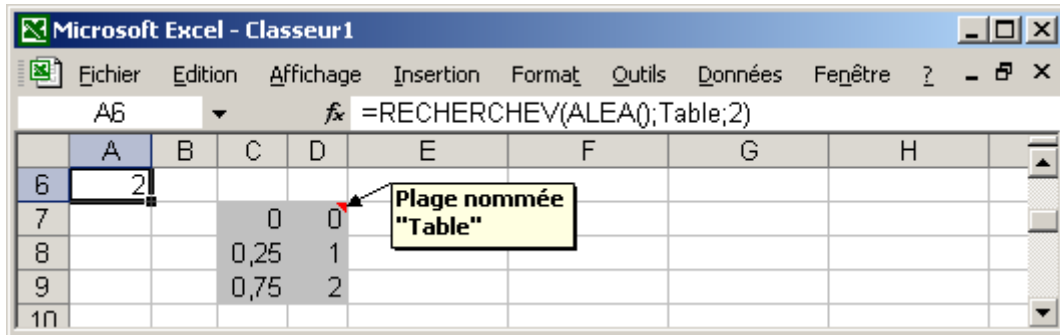
On fait 2 lancers de pièce et on note le nombre de fois où on a retourné le côté face. La variable correspondante est distribuée de la manière suivante :

| x | Proba | Proba cumulée $F(x)$ |
|---|-------|-------------------------|
| 0 | 0,25 | 0,25 |
| 1 | 0,5 | 0,75 |
| 2 | 0,25 | 1 |



La réciproque de la fonction de répartition fait correspondre la valeur $x = 0$ aux valeurs de probabilité plus petites que 0,25, la valeur $x = 1$ aux valeurs comprises entre 0,25 et 0,75 et la valeur $x = 2$ aux valeurs de probabilité supérieures à 0,75.

La correspondance peut être faite sur Excel avec la fonction **RECHERCHEV**. (voir la fiche « Fonction »)



Vérifiez que les valeurs de la variable apparaissent avec des fréquences proches des probabilités.

2. Le cas de la loi discrète uniforme

On veut simuler le point marqué par un dé. Pour faire correspondre la valeur 1 aux valeurs de la fonction **ALEA** comprises entre 0 et 1/6, la valeur 2 aux valeurs entre 1/6 et 2/6, etc., il suffit de multiplier le résultat de la fonction **ALEA** par 6, d'éliminer les décimales du résultat et d'ajouter une unité, ce que fait la formule

$$=ENT(ALEA()*6)+1$$

3. Le cas des variables binaires

Reprenons le lancer d'une pièce. Pour simuler le résultat, il suffit de dire qu'on a retourné le côté face si le résultat de la fonction **ALEA** est plus petit que 0,5 :

$$=SI(ALEA())<1/2;"Pile";"Face")$$

De manière générale, on simule la réalisation d'un événement A de probabilité p par la formule

$$=SI(ALEA())<p;"A";"Non A")$$

et la variable indicatrice de l'événement, c'est-à-dire la variable qui prend la valeur 1 si l'événement est réalisé et 0 sinon (loi « de Bernouilli »), par

$$=SI(ALEA())<p;1;0)$$

E. Applications

En appliquant la méthode des fractiles, on simule avec Excel toutes les lois de probabilités usuelles (pour une présentation de ces lois, voir le manuel et l'ouvrage de G. Saporta [2]).

| <i>Loi</i> | <i>Formule</i> |
|---|---|
| discrète uniforme sur les valeurs entières de 1 à n (voir le § D.2 p. 5) | =ENT(ALEA()*n)+1 ou, avec les versions récentes d'Excel, =ALEA.ENTRE.BORNES(1;n) |
| Bernouilli de paramètre p (voir le § D.3 p. 5) | =SI(ALEA()<p;1;0) |
| binomiale de paramètres n et p (voir le § IV.A p. 7) | =CRITERE.LOI.BINOMIALE(n;p;ALEA()) |
| Poisson de paramètre m | On peut utiliser la convergence de la loi binomiale vers la loi de Poisson et reprendre la formule du dessus en remplaçant n par 10^4 et p par m/10^4 |
| hypergéométrique (voir le § IV.C p. 9) | utiliser RECHERCHEV (voir le § D.1 p. 4) à partir des probabilités cumulées calculées avec la fonction LOI.HYPERGEOMETRIQUE |
| géométrique de paramètre p (voir le § IV.B p. 7) | =ARRONDI.SUP(LN(ALEA())/LN(1-p);0) le fractile d'ordre $1-\alpha$ étant l'entier supérieur ou égal à $\ln(\alpha)/\ln(1-p)$ |
| continue uniforme entre 0 et 1 | =ALEA() |
| continue uniforme entre a et b | =ALEA()*(b-a)+a |
| gamma de paramètre r | =LOI.GAMMA.INVERSE(ALEA());r;1) |
| bêta de paramètres n et p entre les bornes a et b | =BETA.INVERSE(ALEA());n;p;a;b) |
| normale standard | =LOI.NORMALE.STANDARD.INVERSE(ALEA()) |
| normale de paramètres m et sigma | =LOI.NORMALE.INVERSE(ALEA());m;sigma) |
| log-normale | =LOI.LOGNORMALE.INVERSE(ALEA());m;sigma) |
| Khi-deux à nu ddl | =KHIDEUX.INVERSE(ALEA());nu) |
| Student à nu ddl | =LOI.STUDENT.INVERSE(ALEA());nu) |
| Fisher à nu1 et nu2 ddl | =INVERSE.LOI.F(ALEA());nu1;nu2) |
| exponentielle de paramètre lambda | =-1/lambda*LN(ALEA()) (voir le chap. « Probabilités et jugement sur échantillon ») |

IV. Méthodes à usage pédagogique

A. La loi binomiale à partir du processus de Bernouilli

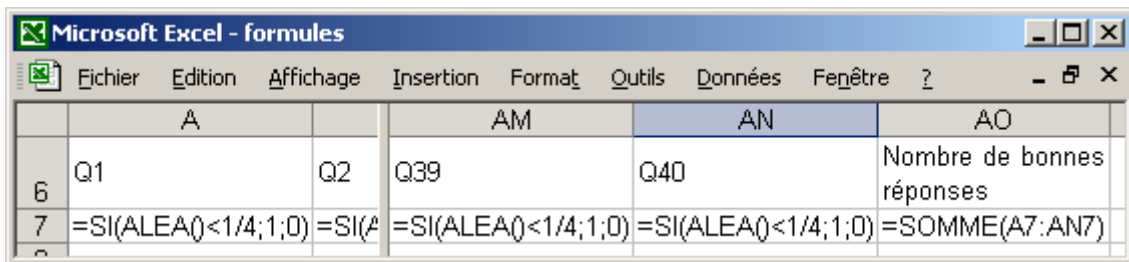
Un examen consiste en une série de 40 questions indépendantes comptant chacune pour le même nombre de points. A chaque question, 4 réponses sont proposées dont une seule est exacte. A-t-on des chances d'avoir plus de la moyenne quand on répond au hasard ?

Le nombre de bonnes réponses suit une loi binomiale, qu'on peut simuler par la formule

=CRITERE.LOI.BINOMIALE(n;p;ALEA())

avec $n = 40$ (nombre de questions) et $p = \frac{1}{4}$ (4 choix possibles par question).

Mais c'est avant tout la somme des indicatrices de succès à chaque question (on note 1 si la réponse à la question est bonne, 0 sinon). C'est le « processus de Bernouilli », qu'on simule par la ligne de formule suivante :



| | A | | AM | AN | AO |
|---|---------------------|---------------------|---------------------|---------------------|---------------------------|
| 6 | Q1 | Q2 | Q39 | Q40 | Nombre de bonnes réponses |
| 7 | =SI(ALEA())<1/4;1;0 | =SI(ALEA())<1/4;1;0 | =SI(ALEA())<1/4;1;0 | =SI(ALEA())<1/4;1;0 | =SOMME(A7:AN7) |

Recopiez les formules sur 1000 lignes. Représentez la distribution du nombre de bonnes réponses par un diagramme en bâtons (voir la fiche « Comment faire »). Comparez la à la distribution de probabilités (calculées avec la fonction **LOI.BINOMIALE**). L'étudiant a-t-il des chances d'avoir coché la bonne réponse à plus de la moitié des questions ?

B. La loi géométrique à partir de sa définition

On tire une bille dans un sac contenant des billes blanches et des billes rouges. Si la bille est blanche, on la remet dans le sac et on tire à nouveau jusqu'à ce qu'on obtienne une bille rouge.

En faisant appel aux probabilités conditionnelles, on montre que le nombre de tirages nécessaires suit la loi de probabilité suivante (voir le chap. « Probabilités et jugement sur échantillon ») :

$$P(X = x) = p(1-p)^{x-1}$$

où p est la proportion de billes rouges dans le sac.

Pour le vérifier, simulons l'épreuve des billes. On dira que la proportion p est égale à 0,5 et on fixera le nombre maximum d'essais à 15.

Repérons le nombre d'essais nécessaires par une indicatrice. Dans la capture d'écran ci-dessous, il a fallu 2 essais.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|-----------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----------------------------|
| 1 | N° de l'essai | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Nombre d'essais nécessaire |
| 2 | Indicatrice de succès | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

Le nombre d'essais est calculé en faisant la somme des produits des valeurs de la ligne des numéros et de celle des indicatrices

| | A | B | C | D | E | J | K | L | M | N | O | P | Q | R |
|---|-----------------------|---|---|---|---|---|----|----|----|----|----|----|----------------------------|---|
| 1 | N° de l'essai | 1 | 2 | 3 | 4 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Nombre d'essais nécessaire | |
| 2 | Indicatrice de succès | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | |

Au premier essai, l'indicatrice de succès peut être simulée avec la formule suivante (voir p. 5)

$$=SI(ALEA()<p;1;0)$$

Pour les essais suivants, de deux choses l'une :

- ❖ Soit l'événement n'a pas été réalisé. Dans ce cas, la probabilité p de réussite à chaque tirage restant constante, on reprend la formule du dessus
- ❖ Soit l'événement a déjà été réalisé. Dans ce cas, le tirage n'a pas lieu.

Pour voir si l'événement a été réalisé ou non aux tentatives précédentes, il suffit de faire la somme des indicatrices.

| | A | B | C | D | E | J | K | L | M | N | O | P | Q | R |
|---|-----------------------|---|---|---|---|---|----|----|----|----|----|----|----------------------------|---|
| 1 | N° de l'essai | 1 | 2 | 3 | 4 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Nombre d'essais nécessaire | |
| 2 | Indicatrice de succès | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | |

Ces formules sont reprises dans le document « Billes ». Téléchargez le. Recopiez les formules sur 1000 lignes. Représentez la distribution des valeurs observées et vérifiez qu'elle correspond à la loi définie plus haut.

C. La loi hypergéométrique à partir d'un tirage sans remise

On prend 5 œufs dans un panier de 10 œufs. Si 3 œufs sur les 10 sont pourris, combien en a-t-on de pourris sur les 5 ?

Les initiés ont reconnu la loi « du tirage sans remise », encore appelée « loi géométrique ».

Pour simuler l'exemple, téléchargez le document « Omelette ».

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|-----------------------------|------|------|------|------|------|------|------|------|------|------|--|
| 1 | N° de l'œuf | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Nombre d'œufs pourris dans l'échantillon |
| 2 | Etat : 1 si pourri | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3 | Aléa | 0,06 | 0,61 | 0,71 | 0,29 | 0,46 | 0,14 | 0,03 | 0,01 | 0,69 | 0,76 | |
| 4 | Présence dans l'échantillon | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |

On affecte à chacun des œufs du panier une valeur aléatoire (fonction **ALEA**) et on choisit les 5 premiers

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|-----------------------------|------|------|------|------|------|------|------|------|------|------|--|
| 1 | N° de l'œuf | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Nombre d'œufs pourris dans l'échantillon |
| 2 | Etat : 1 si pourri | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3 | Aléa | 0,06 | 0,61 | 0,71 | 0,29 | 0,46 | 0,14 | 0,03 | 0,01 | 0,69 | 0,76 | |
| 4 | Présence dans l'échantillon | 1;0) | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |

Les 5 premières valeurs sont **0,01** (œuf n°8), **0,03** (n°7), **0,06** (n°1), **0,14** (n°6) et **0,29** (n°4).

Par conséquent, on casse les œufs n° 8, 7, 1, 6 et 4, parmi lesquels se trouve un œuf pourri : le 1.

Le nombre d'œufs pourris est calculé en faisant la somme des produits (fonction **SOMMEPROD**) des indicatrices de l'état et de celles de la présence dans l'échantillon.

Simulez 1000 tirages en suivant la petite astuce de la feuille Recopie. Vérifiez que les fréquences sont proches des probabilités données par la fonction **LOI.HYPERGEOMETRIQUE**.

D. Autres lois

- ❖ La loi de Laplace-Gauss à partir du théorème de la limite centrée (voir le chap. « Probabilités et jugement sur échantillon »)
- ❖ La loi du Khi-deux à partir de la somme des carrés de lois normales centrées et réduites (idem)
- ❖ Etc.

V. Une application en analyse du risque

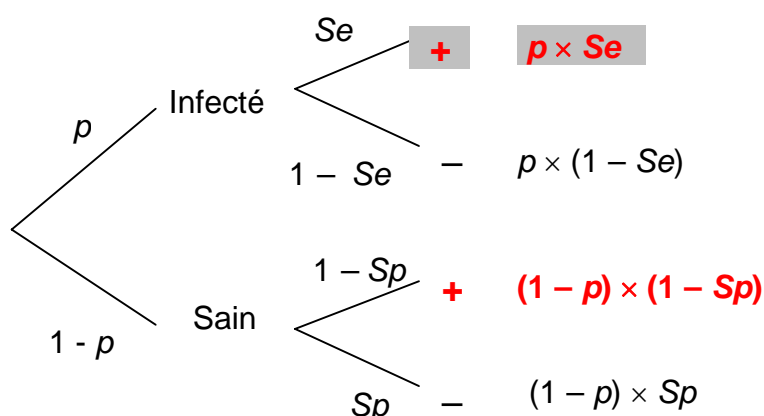
A. Le problème

Un animal provient d'une zone dans laquelle la prévalence d'une maladie est égale à 1%. L'animal subit un test sérologique de sensibilité $Se = 0,9$ et de spécificité $Sp = 0,99$. Il est positif à ce test. Quelle est la probabilité que l'animal soit infecté ? (tiré de l'article de R. Pouillot et M. Sanaa [3])

Le lecteur n'étant pas nécessairement initié au jargon de l'épidémiologie, commençons par définir les termes « prévalence », « sensibilité » et « spécificité ».

- ❖ Le taux de prévalence est la proportion d'individus infectés dans la population étudiée. C'est donc la probabilité p qu'un individu choisi de manière aléatoire dans la population soit infecté.
- ❖ La sensibilité est la probabilité qu'un individu infecté soumis au test présente un résultat de test positif.
- ❖ La spécificité est la probabilité qu'un individu sain soumis au test présente un résultat de test négatif.

Ces définitions étant posées, construisons l'arbre des probabilités.



Le résultat du test étant positif, on est nécessairement sur la première branche terminale ou sur la 3^e (**branches en rouge**) et, parmi ces branches, c'est la première (**rouge et fond grisé**) qui correspond à la situation où l'animal est infecté.

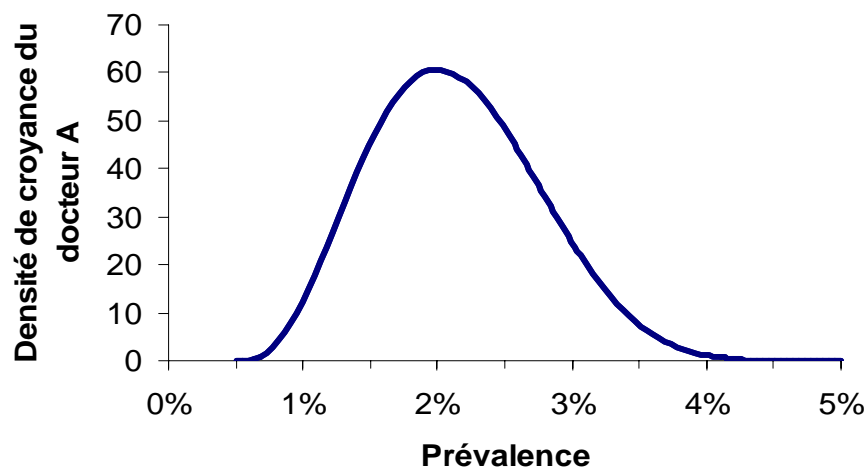
La probabilité que l'animal soit infecté sachant que le résultat du test est positif est donc la probabilité de la branche **1** divisée par la somme des probabilités des branches **1** et **3**, soit

$$\frac{pSe}{pSe + (1-p)(1-Sp)} = \frac{0,01 \times 0,9}{0,01 \times 0,9 + (1-0,01)(1-0,99)} = 0,48$$

Pour le moment, nous n'avons pas fait usage de la simulation parce que nous avons pris un cas où les valeurs des paramètres sont connues.

Dans la réalité, ce n'est malheureusement pas si simple. Ainsi, bien souvent on ne fait qu'apprécier la prévalence à travers des opinions d'experts.

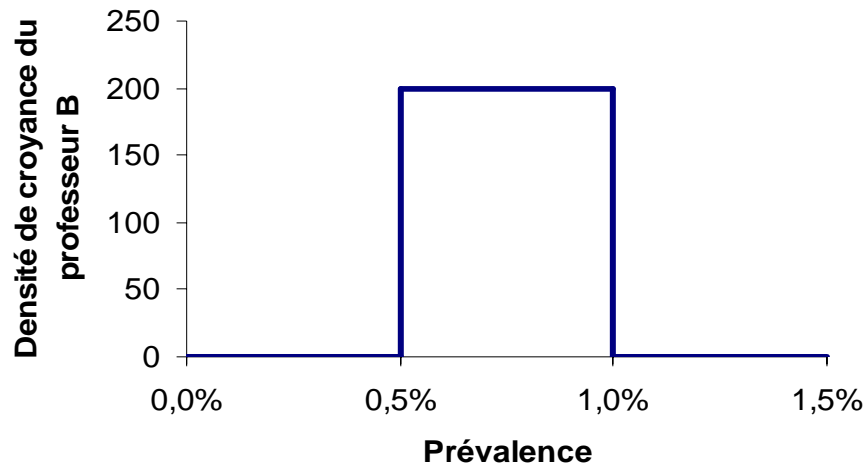
Le premier expert consulté, le docteur A, situe la prévalence entre 0,5% et 4%, au maximum 5%, avec une préférence autour de 2%. Modélisée par une loi bêta, loi qui se prête assez bien à la chose, l'opinion du professeur A se présente de la manière suivante



Téléchargez le document Bêta. Vous y trouverez la densité de la loi bêta, Excel ne proposant pas de fonction permettant de la calculer.

L'opinion du docteur A correspond à la loi bêta de paramètres $a = 4$ et $b = 7$ entre les bornes 0,5% et 5%.

Mais le docteur A n'est pas le seul expert à avoir autorité sur le sujet. Il faut aussi prendre en compte l'opinion du professeur B. Ce dernier a un avis différent : il situe la prévalence entre 0,5% et 1%, ceci sans préférence particulière.

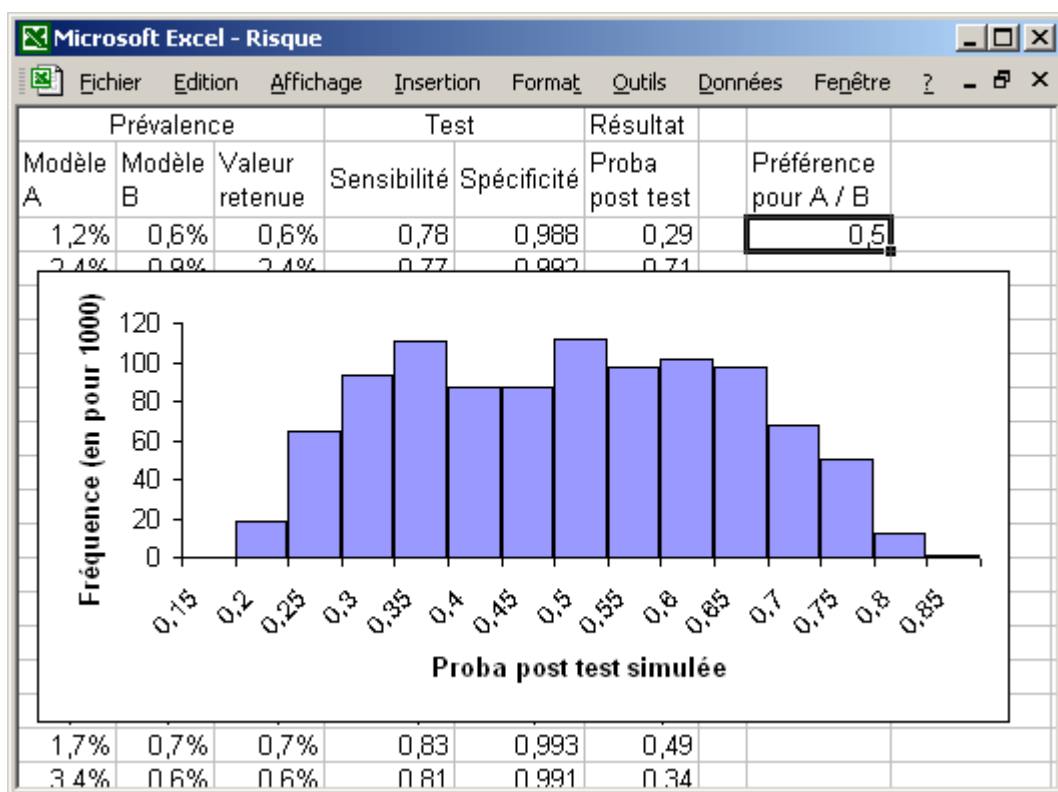


On modélise l'opinion du professeur B par la loi uniforme entre 0,5% et 1%.

Comment gérer cette situation, sachant que, pour compliquer le problème, la sensibilité et la spécificité du test sont données sous la forme d'une fourchette : entre 0,85 et 0,95 pour la sensibilité et entre 0,985 et 0,995 pour la spécificité ?

B. Evaluation du risque par simulation

Téléchargez le document Risque.



La prévalence est simulés selon chacun des modèles.

- ❖ La prévalence selon le modèle du docteur A suit la loi bêta de paramètres 4 et 7 entre les bornes 0,5% et 5%, qu'on simule par la formule suivante (voir p. 5)

=BETA.INVERSE(ALEA();4;7;0,005;0,05)

- ❖ La prévalence selon le modèle du professeur B suit la loi uniforme entre les bornes 0,5% et 1%, simulée par la formule

= ALEA()*(0,01-0,005)+0,005

Le choix du modèle se fait selon un poids (nommé **poids_AvsB**) traduisant la préférence pour l'un ou l'autre expert. Si on accorde la même crédibilité à chacun des experts, ce poids est égal à 0,5.

On prend l'une ou l'autre des prévalences issues de chacun des modèles avec la formule utilisée pour les variables binaires (p. 5)

=SI(ALEA()<poids_AvsB;choix de A;choix de B)

Prenez connaissance des formules.

Faites varier la préférence pour l'un ou l'autre expert et observez la distribution des probabilités post-test.

VI. Références

- [1] Le manuel du groupe « Le Cercle d'Excel'ense » : Morineau A., Chatelin Y.-M. (Coordinateurs) – L'analyse statistique des données. Apprendre, comprendre et réaliser avec Excel. Editions Ellipses, 2005
- [2] Saporta G. – Probabilités, Analyse des Données et Statistique. Editions Technip, 1990
- [3] Pouillot R., Sanaa M. – Bases probabilistes et statistiques nécessaires à l'appréciation du risque. *In* Epidémiol. et santé anim., 2002, 41, pp. 67-83