CARACTERISATION D'UNE LOI DE DISTRIBUTION D'UNE VARIABLE ENTRANT DANS UN MODELE D'ANALYSE DE RISQUE PROBABILISTE

***R. Pouillot*[1]**

RESUME : *Cet article développe quelques aspects techniques pour la caractérisation d'une loi de distribution d'une variable entrant dans un modèle d'appréciation quantitative des risques liés à l'importation d'animaux vivants ou de denrées alimentaires. Quatre principales techniques sont présentées, selon les données disponibles et le type de variable à modéliser. Des exemples pratiques sont proposés, ainsi que les caractéristiques des principales distributions paramétriques utilisées dans ce domaine.*

SUMMARY : *This paper presents some technical aspects to specify distributions from variables included in a quantitative risk assessment model linked to importation of live animals or food. Four major methods are described, according to the available data and the variable to model. Practical examples are proposed, and the main characteristics of the usual distribution used in this field of interest are recalled.*

Afin de réaliser une appréciation quantitative de risque (AQR) probabiliste (encore appelée AQR « stochastique »), il est nécessaire de définir, pour chacune des variables aléatoires entrant dans le modèle, une loi de distribution représentant les différentes probabilités associées à toutes les réalisations possibles de la variable [Pouillot *et al.*, 2002, ce numéro].

Cet article a pour objectif de proposer quelques points techniques permettant de préciser ces lois de distribution, selon le type de variable ou selon le type de données disponibles.

Il n'est pas possible, dans cet article, de présenter l'ensemble des théories statistiques et probabilistes nécessaires à la réalisation de cet objectif. La connaissance des bases présentées dans l'article de Pouillot et Sanaa [2002, ce numéro] est un strict minimum.

Afin de modéliser une variable aléatoire par une loi de distribution, il est nécessaire de choisir :

- la forme de la loi de distribution suivie par la variable : il existe une infinité de lois de distribution, toute fonction dont l'intégrale est une fonction monotone, croissante et dont l'intégrale sur l'intervalle [-∞ ; +∞] est égale à 1 étant une loi de distribution. Cependant, certaines lois de distribution sont particulières et ont été nommées (exemple : loi normale, loi binomiale, loi gamma, etc.) ;
- les paramètres de la distribution : selon la forme de la loi de distribution choisie, le nombre de paramètres nécessaires pour définir complètement cette loi varie de un (loi de Poisson) à, en théorie, l'infini (loi personnalisée). Il est souvent de deux.

[1] Afssa, 27-31 avenue du Général Leclerc, BP 19, 94701 Maisons-Alfort cedex, France

Une fois la **forme** de la distribution et ses **paramètres** définis, il est possible de connaître la densité de la variable aléatoire pour l'ensemble de son domaine de définition.

Exemple : une loi normale de moyenne $\mu = 0$ et d'écart type $\sigma = 1$ est parfaitement définie sur [-∞ ;+ ∞]. Il est possible de connaître la densité de cette loi pour toute valeur de la variable.

Pour modéliser la distribution d'une variable aléatoire, il est possible de considérer, pour schématiser, quatre éventualités selon le type de la variable et les données disponibles :

1. si le type d'expérience permettant d'obtenir la variable est un type d'expérience classique, que nous nommerons [selon Vose, 2000] « processus stochastique », la théorie probabiliste nous donnera la forme et les paramètres de la distribution ;
2. dans certains cas, la théorie statistique nous permettra de définir la loi de distribution suivie par la variable et ses paramètres associés ;
3. si l'on dispose d'un grand nombre de données observées, il existe alors deux solutions : l'utilisation d'une loi empirique ou l'ajustement des données par une loi de distribution ;
4. enfin, il est possible, le cas échéant, de définir la loi de distribution d'un paramètre à partir de dires d'experts.

Le plan proposé reprend ces quatre possibilités. Les principales lois de distribution utiles en AQR probabiliste liée à l'importation d'animaux ou de denrées animales sont présentées dans cet article sous forme d'encadrés. La plupart des densités des lois de distributions présentées dans cet article peuvent être calculées à l'aide du logiciel Excel® (Microsoft, Corp.) : nous présenterons les formules Excel® permettant de calculer ces densités. Nous présenterons également les fonctions du logiciel @Risk® (Palisade, Corp), permettant de simuler des valeurs issues des lois de distribution proposées.

I – PREMIER CAS

Le type d'expérience est un processus connu : les processus stochastiques

Certaines variables aléatoires sont particulières, car elles correspondent à la réalisation d'un type d'expérience particulier, rencontré dans de nombreux domaines et généralisable : on appellera ces types d'expériences « processus stochastiques ».

Certaines généralisations semblent assez évidentes. Par exemple, la forme de la loi de probabilité du nombre de boules noires obtenues par tirage au sort d'une urne contenant des boules noires et blanches est la même que celle du nombre d'animaux infectés tirés au sort d'une population contenant des animaux indemnes et infectés : dans les deux cas, on a deux caractéristiques dans la population, on tire au sort des individus de cette population et l'on compte le nombre d'individus présentant une des caractéristiques dans cet échantillon ; ce processus est appelé « processus hypergéométrique ». D'autres généra-lisations sont moins évidentes : par exemple, il semble *a priori* difficile de rapprocher la loi de distribution du nombre de bactéries présentes dans un certain volume de lait, la loi de distribution du nombre de foyers d'une maladie sporadique par an, ou la loi de distribution du nombre de pièces défectueuses sortant d'une machine de production par heure. Or il est possible, sous certaines conditions, de considérer ces trois types d'évènements comme des réalisations d'un même processus particulier dit « processus de Poisson ».

On présentera ici les notions élémentaires de trois processus stochastiques principaux : le processus binomial, le processus hypergéométrique et le processus de Poisson. Une bonne compréhension de ces notions permet de résoudre la plupart des problèmes de caractérisation de variables aléatoires en appréciation quantitative des risques liés à l'importation.

1. LE PROCESSUS BINOMIAL

Expérience type : On dispose d'une population (d'animaux, de cheptels, de lots d'aliments, de résultats d'analyses, etc.). Dans cette population, une proportion *p* d'entités possède une caractéristique particulière, la différenciant des autres entités. On tire au sort **avec remise** de cette population un échantillon de taille *n* entités et l'on compte le nombre d'entités présentant la caractéristique d'intérêt dans cet échantillon : ce nombre d'entités sera noté *s* (comme « succès »). « Avec remise » signifie que l'on replace, après chaque tirage au sort, l'entité dans la population : une entité peut donc être tirée au sort plusieurs fois.

Un processus binomial a cinq caractéristiques ; si l'on est face à un exemple non proposé dans ce chapitre, il suffira de se demander si l'expérience que l'on désire simuler répond à ces cinq principes :

1. l'expérience consiste à réaliser *n* essais identiques (un essai est le tirage au sort d'un individu de la population) ;
2. la conclusion de chaque essai se résume à deux possibilités (que l'on appellera « succès » et « échec ») ;
3. la probabilité de succès à chaque essai est constante durant toute l'expérience, et est égale à *p* ;
4. les essais sont réalisés de façon indépendante : le résultat d'un essai n'est pas influencé par les résultats précédents ou suivants ;
5. on s'intéresse à *s* le nombre de succès dans la population.

Les troisième et quatrième conditions sont violées si l'on effectue un tirage au sort sans remise (sans remettre l'entité tirée au sort dans la population). Prenons l'exemple du tirage au sort de cinq animaux dans un lot de 10 animaux parmi lesquels deux animaux sont infectés : à la suite du premier tirage au sort, si on a choisi un animal infecté, la probabilité de tirer un second animal infecté n'est plus de 2/10 mais de 1/9 : la probabilité varie au cours de l'expérimentation (violation de la condition 3) et le résultat des tirages au sort est influencé par le résultat des tirages au sort précédents (violation de la condition 4).

En pratique, on n'effectue que très rarement un tirage au sort avec remise. Il serait alors nécessaire d'utiliser un autre processus, dit « processus hypergéométrique » (*cf.* infra). Cependant, le processus binomial est très souvent utilisé dans le domaine de l'appréciation des risques à l'importation car on admet qu'il peut être employé dès que moins de 10% de la population totale est échantillonnée. Alors, la variation de la probabilité de succès dans la population suite aux premiers tirages au sort est considérée comme négligeable.

Prenons l'exemple du tirage au sort de n = cinq animaux d'une population de 1 000 animaux renfermant 10 animaux infectés (p = 10/1000 = 1%). Ces cinq animaux représentent moins de 10% de la population totale. On peut alors considérer que l'on est face à un processus binomial. Ainsi, si l'on tire un animal au sort, et que cet animal est infecté, la probabilité de succès lors du second tirage au sort sera de 9/999 = 0,9% $\approx$ 1%.

Lorsque le processus d'obtention d'une variable aléatoire est binomial, la connaissance de deux paramètres parmi *n*, *p*, et *s* permet de faire des inférences sur le troisième paramètre (tableau I). Les paragraphes suivants présentent ces relations. On notera que la connaissance de la taille de la population globale n'est pas nécessaire dans un processus binomial.

TABLEAU I

Relations entre les trois paramètres d'un processus binomial

Paramètres connus	Loi du troisième paramètre
n, *p*	*s* ~ *binomiale*
s, *p*	*n* ~ *binomiale négative*
n, *s*	*p* ~ *bêta**

* : une des lois possibles (cf. infra)

1.1. NOMBRE DE SUCCES DANS L'ECHANTILLON : LA LOI BINOMIALE

Dans un processus binomial, on connaît n, la taille de l'échantillon et p, la probabilité de « succès » dans la population. On cherche à modéliser la variable s, nombre de « succès » dans l'échantillon. s suit une loi de distribution particulière dite « loi binomiale ». On notera

$$s \sim Binom(n, p),$$

le symbole « ~ » se lisant : « suit une loi de distribution ». En pratique, cette loi est donc utilisée lors de :

1) tout tirage au sort dans une population avec remise ;
2) tout tirage au sort dans une population « infinie » : cette notion de population infinie devant être considérée comme « de très grande taille au regard de l'échantillon ». En pratique, on retiendra la loi binomiale lorsque l'échantillon représente moins de 10% de la population totale. **Exemple** : on tire au sort un échantillon limité d'animaux parmi les animaux d'un pays. Le nombre d'animaux présentant une caractéristique particulière dans notre échantillon (état d'infection, sexe) suit une loi binomiale ;
3) toutes réalisations d'un événement de probabilité constante. **Exemple** : on réalise un test de diagnostic sur un certain nombre d'animaux infectés. Le nombre d'animaux détectés positifs suit une loi binomiale.

L'encadré A présente les caractéristiques de cette loi de distribution. On notera le caractère asymétrique de la distribution lorsque p est faible.

PROPRIETES UTILES EN AQR

1) La somme de deux variables aléatoires binomiales **indépendantes** et **de même paramètre** p est une variable aléatoire binomiale. Si $X_1 \sim Binom(n_1, p)$ et $X_2 \sim Binom(n_2, p)$, alors $X_1 + X_2 \sim Binom(n_1 + n_2, p)$. **Exemples** : la somme de « succès », si l'on choisit cinq animaux dans deux lots de grande taille et si la prévalence p est identique dans chacun des lots, suit une binomiale de paramètres $n = 5 + 5 = 10$ et p. La somme de « succès », si l'on choisit cinq animaux dans un lot de grande taille 10 années de suite et si la prévalence p est constante durant ces 10 années, suit une loi binomiale de paramètres 50 et p.
2) Si le nombre de succès suit une loi binomiale de paramètre n et p, le nombre d'échecs suit une loi binomiale de paramètres n et $(1 - p)$.
3) Si le produit np et le produit $n(1 - p)$ sont grands (en pratique > 5), la loi binomiale $Binom(n, p)$ tend vers la loi normale $N(np, \sqrt{np(1-p)})$ (*cf.* infra). **Exemple** : le nombre d'alevins « femelles » parmi 500 alevins tirés au sort d'une grande population suit approximativement une loi normale $N(250, 11{,}2)$ (si $p = 50\%$.). On a en effet $np = n(1 - p) = 250$ ce qui est supérieur à 5 et $\sqrt{np(1-p)} = \sqrt{500 \times 0{,}5 \times (1-0{,}5)} = 11{,}2$.
4) Si n est grand (en pratique > 40) et le produit np est petit (en pratique < 5), la loi binomiale $Binom(n, p)$ tend vers la loi de Poisson $Poisson(np)$ (*cf.* infra). **Exemple** : le nombre d'animaux infectés dans un lot de 100 animaux issu d'une population où la prévalence de la maladie est de 0,1% suit approximativement une loi de Poisson $Poisson(0{,}1)$.

Exemple classique dans notre domaine : La prévalence individuelle réelle d'une maladie dans un pays est égale à 5%. On choisit au hasard 100 animaux dans ce pays. Le nombre d'animaux infectés dans notre échantillon suivra une loi binomiale de paramètre $n = 100$ et $p = 5\%$. Nota : la condition de constance de la probabilité p dans la population globale peut être violée en cas de tirages au sort réalisés durant une grande période de temps. Par exemple, si l'on tire au sort un cheptel par jour pendant trois mois durant une épizootie de fièvre aphteuse, la prévalence dans la population globale ne peut être considérée comme constante. Une des conditions nécessaires pour satisfaire la constance de la probabilité p peut être une relative « rapidité » de l'échantillonnage.

Encadré A : Loi binomiale

Ecriture $x \sim Binom(n, p)$

pour x : entier $n \geq x \geq 0$

avec n : nombre d'essais (entier ≥ 0)

et p : probabilité dans la population ($\in [0 ; 1]$)

Espérance np

Variance $np(1 - p)$

Densité $$f(x) = C_n^x p^x (1-p)^x$$

Fonction de répartition $$F(x) = \sum_{y=0}^{x} C_n^y p^y (1-p)^y$$

Estimation des paramètres (méthode des moments)

$$\hat{p} = \frac{r}{n}$$

avec r : réalisation d'un nombre de succès parmi n essais.

Propriétés

- si $X \sim Binom(n, p)$ alors son contraire $\overline{X} \sim Binom(n, (1 - p))$;
- la loi binomiale tend vers la loi normale $N(np, \sqrt{n(1-p)})$ lorsque np et $n(1 - p)$ sont grands (en pratique np et $n(1 - p) > 5$) ;
- la loi binomiale tend vers la loi de Poisson $Poisson(np)$ lorsque n est grand et np est petit (en pratique $n > 40$ et $np < 5$) ;
- Si $X_1 \sim Binom(n_1, p)$, $X_2 \sim Binom(n_2, p)$ et X_1 et X_2 sont indépendants, alors $X_1 + X_2 \sim Binom(n_1 + n_2, p)$.

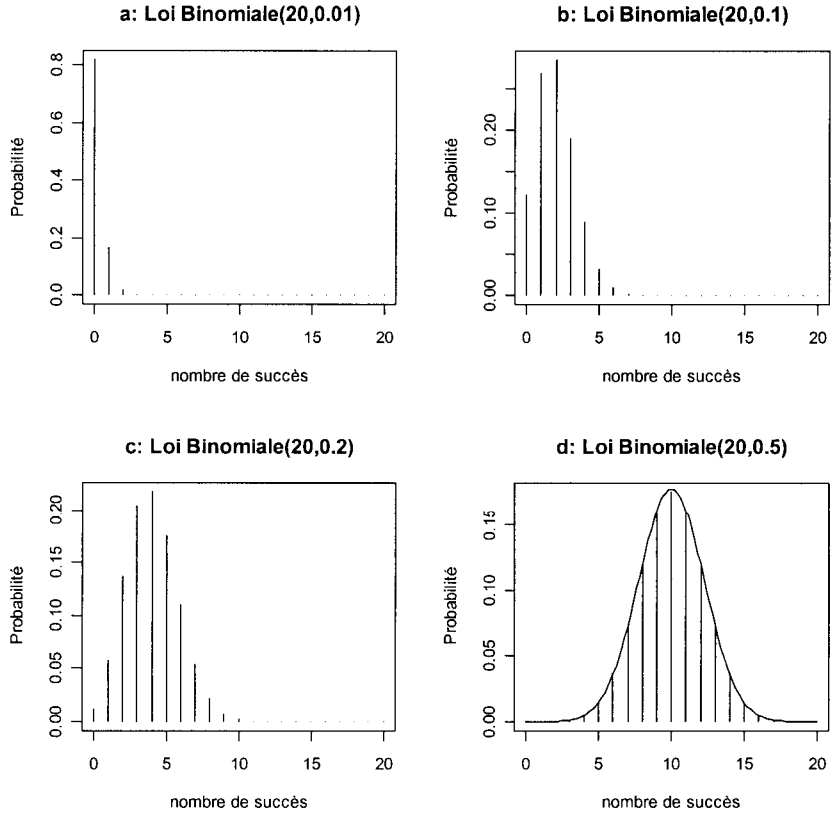
Ecriture Excel©

densité = LOI.BINOMIALE(x, n, p, 0)

fonction de répartition = LOI.BINOMIALE(x, n, p, 1)

Ecriture @Risk© =RISKBINOMIAL(n, p)

Exemples : dans la figure d), la courbe représente l'approximation de la loi binomiale par la loi normale.



Exemple 1 : La probabilité annuelle d'introduction d'une maladie sur un territoire est estimée par $p = 0{,}1$. On considère que cette probabilité est constante d'année en année. On importe tous les ans durant 10 ans des animaux de ce pays. Le nombre d'années voyant l'introduction de la maladie suit une loi binomiale de paramètre $n = 10$ et $p = 0{,}1$.

Exemple 2 : Si l'on teste n animaux infectés avec un test de sensibilité *Se*, le nombre d'animaux infectés positifs au test suit une loi binomiale de paramètres n et *Se*. *Justification* : parmi la population des animaux infectés, un certain nombre présente la caractéristique « positif au test », les autres ayant la caractéristique « négatif au test » (selon le niveau d'anticorps circulant, par exemple). Compter le nombre d'animaux infectés positifs au test parmi les n animaux revient à tirer au sort dans la population « globale » d'animaux infectés un certain nombre d'animaux « positifs au test ».

1.1. Nombre d'essais necessaires pour obtenir un nombre connu de succes : la loi binomiale negative

Dans un processus binomial, on connaît s, le nombre de succès dans l'échantillon et p, la proportion de « succès » dans la population totale. On cherche à connaître la taille de notre échantillon (cette question peut paraître étrange au premier abord, mais les exemples ci-dessous illustreront le propos dans notre domaine d'intérêt).

La modélisation de n utilise la loi binomiale négative. Cette loi nécessite deux paramètres : s et p. Elle permet de modéliser le nombre **d'échecs** observés dans un processus binomial avant d'obtenir s succès, si la probabilité de succès dans la population est p.

Deux cas doivent être distingués :

1. le dernier tirage au sort est un succès : on arrête de tirer au sort dans la population **dès que l'on a obtenu le $s^{\text{ième}}$ succès**. On a alors :

   $n \sim s + BinNeg(s, p)$

   où $BinNeg(s, p)$ est la loi binomiale négative de paramètre s et p[2]

2. **on ne sait pas si le dernier tirage au sort est un succès ou un échec** : on a obtenu s succès, on cherche à connaître le nombre d'essais qui ont été réalisés, mais on ne sait pas si le dernier essai observé était un échec ou un succès. En utilisant le théorème de Bayes[3] (démonstration non proposée, *cf.* Vose, 2000), on a :

   $n \sim s + BinNeg(s + 1, p)$

   donc

   si $s \sim Binom(n, p)$:

   $n \sim s + BinNeg(s, p)$ si le dernier essai est un succès ;

   $n \sim s + BinNeg(s + 1, p)$ si l'on ne sait pas si le dernier essai est un succès ou un échec.

Les caractéristiques de la loi binomiale négative sont présentées dans l'encadré B.

Applications

1. On cherche à appréhender le nombre d'années sans importation d'animaux infectés si la probabilité annuelle d'importation d'animaux infectés est égale à p et est constante.
2. On teste des animaux dans une grande population où la prévalence apparente de la maladie est p : on veut savoir combien d'animaux il faudra tester avant d'avoir s animaux positifs.
3. On cherche à modéliser le nombre d'animaux **réellement** infectés n dans une population, après application d'un test de sensibilité *Se* connu et de spécificité 1.

Exemple 1 : On importe des animaux d'un pays. Le risque annuel d'introduction d'une maladie, considéré comme constant, est de $p = 1\%$. Le nombre d'années sans introduction de la maladie suit une loi binomiale négative de paramètre $s = 1$ (on arrête le processus au premier « succès »), $p = 0{,}01$. Le nombre d'années n avant qu'un problème survienne peut être modélisé selon $n \sim BinNeg(1, 0{,}01) + 1$.

[2] La loi binomiale négative nous permet d'évaluer le nombre d'échecs. Pour obtenir le nombre d'essais n, il faut ajouter au nombre d'échecs modélisés par la loi binomiale négative le nombre de succès observés s.

[3] Avec une distribution *a priori* uniforme du nombre d'échecs.

Encadré B : Loi binomiale négative

Ecriture $x \sim BinNeg(s, p)$

pour x : entier ≥ 0
avec s : nombre de succès (entier > 0)
et p : probabilité de succès dans la population ($\in [0 ; 1]$)

Espérance

$$\frac{s(1-p)}{p}$$

Variance

$$\frac{s(1-p)}{p^2}$$

Densité

$$f(x) = C_{s+x-1}^{x} p^s (1-p)^x$$

Fonction de répartition

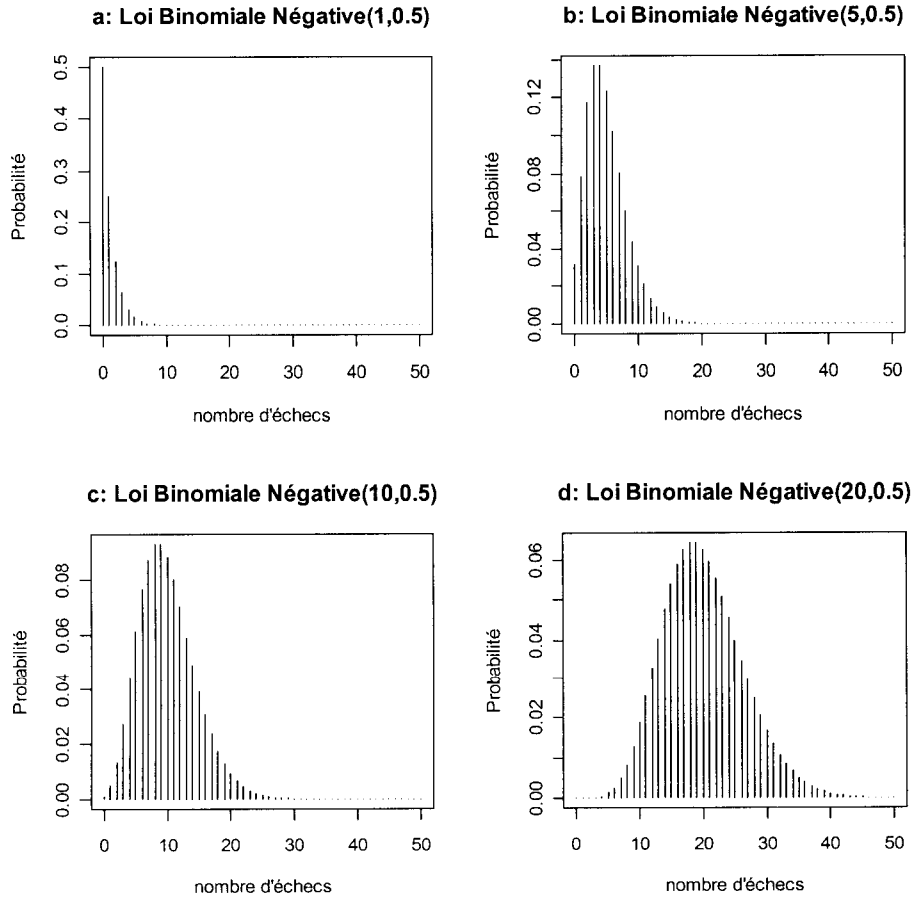
$$F(x) = \sum_{y=0}^{x} C_{s+y-1}^{y} p^s (1-p)^y$$

Estimation des paramètres (méthode des moments)

$$\hat{p} = \frac{s}{n}$$

avec s : réalisation d'un nombre de succès parmi n essais.

Ecriture Excel©
densité =LOI.BINOMIALE.NEG(x, s, p)
Ecriture @Risk© =RISKNEGBIN(s, p)
Exemple



Exemple 2 : Si, dans la situation précédente, on décide d'arrêter l'importation en provenance de ce pays dès lors que l'on aura obtenu trois introductions de maladie, le nombre d'années d'importation suivra une loi *BinNeg*(3, 0,01) + 3.

Exemple 3 : On applique un test de sensibilité *Se* et de spécificité 1 à une population de statut infectieux individuel inconnu. On obtient un certain nombre *s* d'animaux à réponse positive au test, et l'on cherche *n*, le nombre total d'animaux réellement infectés dans la population. On peut présenter ce processus comme suit : chaque animal réellement infecté a une probabilité *Se* d'être détecté positif. On dispose du nombre d'animaux détectés positifs dans la population *s*, mais on cherche le nombre d'essais ayant permis d'obtenir nos *s* succès. Le dernier animal testé n'est pas forcément un animal infecté. On a alors :

$n \sim s + BinNeg(s + 1, Se)$

1.3. PROBABILITE DE SUCCES : LA LOI BETA

On dispose des résultats d'un processus binomial : on a obtenu *s* succès parmi *n* échantillons. L'estimation de la probabilité de succès dans la population est $\hat{p} = s / n$. Il existe cependant une certaine incertitude sur le paramètre $\hat{p}$. Il est ainsi intuitif que $\hat{p}$ sera connu avec d'autant plus de précision que *n* sera grand. Il peut être nécessaire dans une AQR de modéliser l'incertitude de cette estimation $\hat{p}$.

Des théories classiques couramment utilisées dans le domaine (fondées sur des hypothèses pragmatiques et sur l'utilisation du théorème de Bayes) permettent d'utiliser la relation :

$p \sim bêta(s + 1, n - s + 1)$

pour modéliser l'incertitude sur le paramètre $\hat{p}$, où *bêta*(*a*, *b*) est une loi bêta de premier paramètre de forme *a* et de second paramètre de forme *b* (*cf.* encadré C)[4].

Exemple 1 : Pour évaluer le taux de prévalence apparente de cheptels infectés dans un pays, on a testé indépendamment après tirage au sort 1 000 cheptels. On a obtenu cinq cheptels à réponse positive. L'incertitude sur le taux de prévalence de cheptels infectés dans la population peut être modélisée selon :

$p \sim bêta(5 + 1, 1000 - 5 + 1)$

Exemple 2 : Pour évaluer les performances d'un test de diagnostic, on a testé 100 animaux infectés de façon certaine, et obtenu 92 animaux à réponse positive. L'incertitude sur l'estimation de la sensibilité du test peut être modélisée selon :

$Se \sim bêta(92 + 1, 100 - 92 + 1)$

1.4. CONCLUSIONS SUR LE PROCESSUS BINOMIAL

Ce processus binomial est utilisé extrêmement fréquemment en appréciation quantitative des risques liés à l'importation d'animaux vivants. En effet, on considère que le nombre d'animaux (de cheptels) infectés dans un échantillon issu d'une grande population suit un processus binomial, ce qui est correct si la taille de l'échantillon n'est pas trop grande et si la probabilité de tirer au sort un cheptel (ou un animal) infecté est constant.

2. LE PROCESSUS HYPERGEOMETRIQUE

Expérience type : On dispose d'une population (d'animaux, de cheptels, de lots d'aliments, de résultats d'analyses, etc.). Dans cette population, on définit *S* le nombre (et non la proportion) d'entités possédant une caractéristique particulière, les différenciant des autres entités.

---

[4] Cette relation suppose que l'on considère *a priori* que la probabilité de succès dans l'échantillon peut varier de 0 à 1 de façon uniforme (toutes les valeurs entre 0 et 1 sont *a priori* équiprobables). On sera prudent sur la validité de cette hypothèse. Si son utilisation n'a généralement que peu de conséquences, elle peut être très forte lorsque *s* est proche de 0 ou proche de *n* ($\hat{p}$ proche de 0 ou de 1).

Encadré C : Loi bêta

Ecriture *x ~ bêta(a, b)*

pour $x \in [0 ; 1]$
avec *a* : premier paramètre de forme (> 0)
et *b* : second paramètre de forme (> 0)

Espérance

$$\frac{a}{a+b}$$

Variance

$$\frac{ab}{(a+b)^2(a+b+1)}$$

Densité

$$\frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}$$

Estimation des paramètres

- si *x* représente une probabilité dans une population,

$$\hat{a} = s+1, \hat{b} = n-s+1$$

en utilisant le théorème de Bayes (cf. texte) avec *s* succès parmi *n* essais ;

- sinon (méthode des moments),

$$\hat{a} = \frac{m(m-m^2-v)}{v}, \hat{b} = \hat{a}\left(\frac{1-m}{m}\right)$$

avec *m* : moyenne observée de réalisations de *x* ;
v : variance observée de réalisations de *x*.

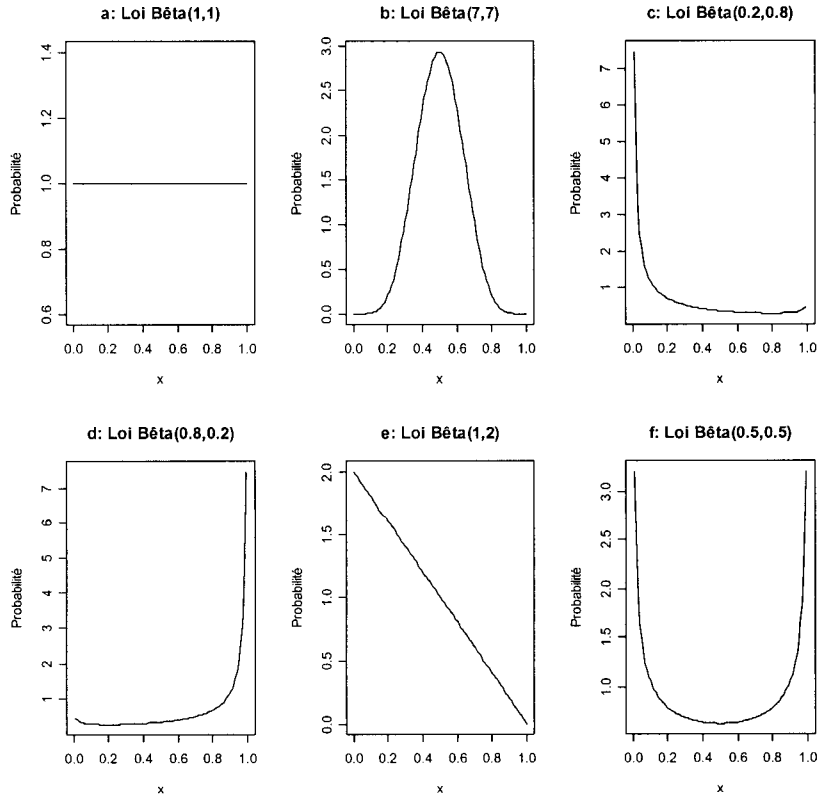
Ecriture Excel©

densité ==LOI.BETA(*x*, *a*, *b*, *linf*, *lsup*)

Nota : l'utilisation de *linf* et *lsup* permet d'étendre le domaine de définition de *x* dans [*linf* ; *lsup*].

Ecriture @Risk© =RISKBETA(*a*, *b*)

Exemples : on observe la diversité de formes possibles de la loi bêta.



On tire au sort sans remise un échantillon de taille *n* individus, et l'on compte le nombre d'entités dans l'échantillon présentant la caractéristique d'intérêt : ce nombre d'entités sera noté *s*. « Sans remise » signifie que l'on ne replace pas, après chaque tirage au sort, l'individu dans la population : un individu ne peut donc être tiré au sort qu'une seule fois. Le processus hypergéométrique est donc similaire au processus binomial, si ce n'est que le tirage au sort est effectué sans remise.

Nota : le processus hypergéométrique devrait être utilisé dans tous les cas de tirage au sort **sans remise**. Cependant, les calculs nécessaires sont faramineux, ce qui justifie l'utilisation de la loi binomiale comme approximation lorsque la taille de l'échantillon est inférieure à 10% de la population totale ($n / N < 0,1$).

Connaissant trois des quatre paramètres parmi *N*, *n*, *S* ou *s*, il est dans certains cas possible de faire des inférences sur le quatrième à l'aide d'une loi paramétrique définie (*cf.* tableau II).

TABLEAU II

Relations entre les quatre paramètres d'un processus hypergéométrique

Paramètres connus	Loi du quatrième paramètre
N, n, S	*s ~ Hypergéométrique*
N, S, s	*n ~ Inverse hypergéométrique* si le dernier essai est un succès Non définie sinon
N, n, s	*S* : Non définie
n, S, s	*N* : Non définie

2.1. NOMBRE DE SUCCES DANS L'ECHANTILLON : LA LOI HYPER-GEOMETRIQUE

On connaît *N*, la taille de la population, *n*, la taille de l'échantillon et *S*, le nombre de succès dans la population. On cherche à modéliser la variable *s*, nombre de succès dans l'échantillon. *s* suit une loi de distribution particulière dite « loi hypergéométrique ». On note :

$$s \sim Hypergéom(N, n, S)$$

En pratique, cette loi est donc utilisée lors de tout tirage au sort sans remise, particulièrement lors de tout tirage au sort dans une population de faible taille au regard de l'échantillon, lorsque le nombre d'animaux présentant le caractère d'intérêt dans la population est connu.

Exemple : si l'on tire au sort 10 animaux dans un cheptel de 50 animaux dont trois sont infectés, le nombre d'animaux infectés dans l'échantillon suit une loi *Hypergéom*(50, 10, 3).

Nota : si l'on dispose de la probabilité *p* de succès dans la population, une estimation de *S* est l'arrondi du produit *pN*. Par exemple, dans une population de taille $N = 967$, si la probabilité d'infection est $p = 0,1$, on utilisera $S = 97$ car $pN = 96,7$.

Les caractéristiques de la loi hypergéométrique sont présentées dans l'encadré D.

2.2. NOMBRE D'ECHANTILLONS A PRELEVER AVANT D'OBTENIR *S* SUCCES : LA LOI HYPERGEOMETRIQUE INVERSE

On connaît *s*, le nombre de succès dans l'échantillon, *N*, la taille de la population et *S*, le nombre de « succès » dans la population totale. On cherche à connaître la taille de l'échantillon *n*. Deux cas doivent être distingués.

1. Le dernier essai est un succès : on arrête de tirer au sort dans la population dès que l'on a obtenu *s* succès. On a alors :

   $$n \sim s + InvHyperGéom(s, S, N)$$

   où *InvHyperGéom (s, S, N)* est la loi hypergéométrique inverse de paramètre *s*, *S* et *N* (*cf.* encadré E).

Encadré D : Loi hypergéométrique

Ecriture $x \sim Hypergéom(N, n, S)$

pour x entier $\max(0, n + S - N) \leq x \leq \min(n, S)$

avec N : taille de la population (entier > 0)

n : taille de l'échantillon (entier $N \geq n \geq 0$)

et S : nombre de succès dans la population (entier $N \geq n \geq 0$)

Espérance

$$\frac{nS}{N}$$

Variance

$$\frac{N-n}{n-1}\frac{nS(N-S)}{N}$$

Densité

$$f(x) = \frac{C_S^x C_{N-S}^{n-x}}{C_N^n}$$

Fonction de répartition

$$F(x) = \sum_{y=0}^{x} \frac{C_S^y C_{N-S}^{n-y}}{C_N^n}$$

Estimation des paramètres (méthode des moments)

$$S = N\frac{r}{n}$$

avec r : nombre de succès observé dans un échantillon de taille n.

Propriétés

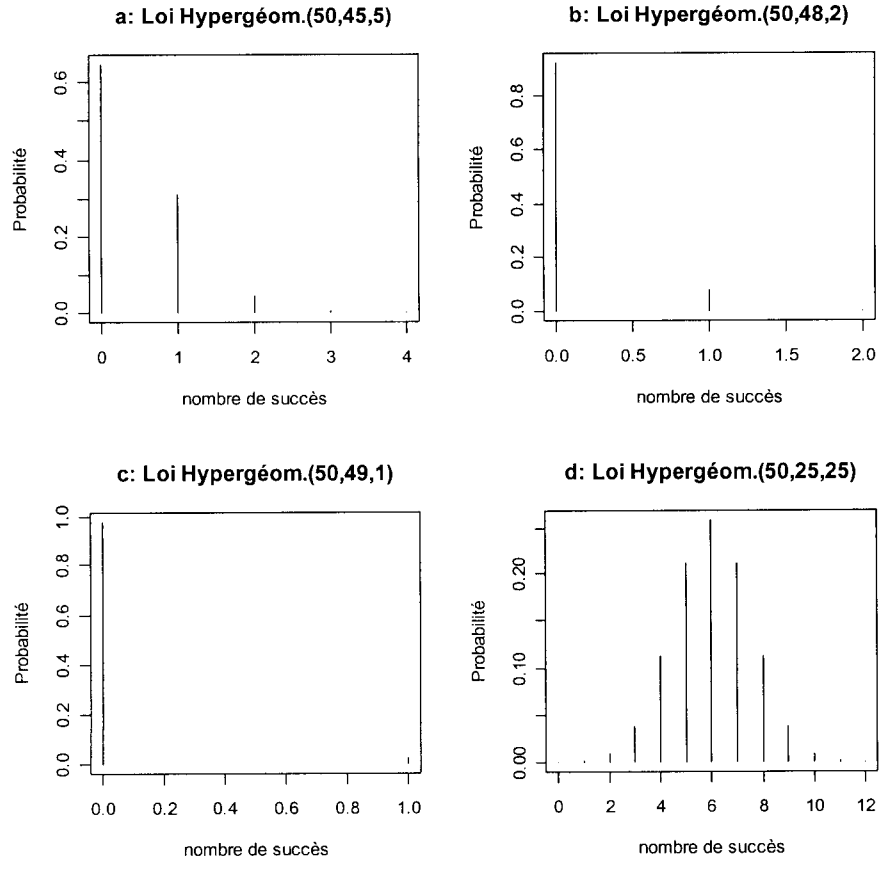
- Si $n << N$ (en pratique, $n/N < 0.1$), $x \sim Binom(n, S/N)$

Ecriture Excel©

densité =LOI.HYPERGEOMETRIQUE(s, N, S, n)

Ecriture @Risk© = RISKHYPERGEO(n, S, N)

Exemples



2. On ne sait pas si le dernier essai est un succès ou un échec. Dans ce cas, la distribution de la taille de l'échantillon prélevé n'a pas de forme simple : on se référera à Vose (2000) pour plus de détails.

Application du cas 1)

On teste des animaux dans une population de faible taille N (exemple : un cheptel) dans laquelle on connaît le nombre d'animaux infectés : on veut savoir combien d'animaux n il faudra tirer au sort avant d'obtenir s animaux infectés.

Exemple : Dans un cheptel de taille $N = 40$, on a $S = 5$ animaux infectés. On désire connaître la distribution du nombre d'animaux à échantillonner n avant d'obtenir le premier animal infecté. On a $n \sim 1 + InvHyperGéom(1, 5, 40)$.

Propriétés utiles en AQR :

1. Si la taille de la population est très grande par rapport au nombre de succès obtenus, la loi inverse hypergéométrique tend vers la loi binomiale négative :

   $X \sim InvHypergéom(s, S, N)$, et $N >> s$ alors $X \sim BinNeg(s, S/N)$

2. Si la taille de la population est très grande par rapport au nombre de succès dans la population, la loi inverse hypergéométrique tend vers la loi gamma (*cf.* infra) :

   $X \sim InvHypergéom(s, S, N)$, et $N >> S$ alors $X \sim gamma(s, S/N)$

2.3. Taille de la population *N*, nombre d'individus présentant la caracteristique d'interet dans la population *S*

Pour modéliser la taille de la population N connaissant s, S, et n et pour modéliser le nombre d'individus présentant la caractéristique d'intérêt dans la population S connaissant N, s, et n, il n'existe pas de loi de distribution simple. Le lecteur confronté à ce problème pourra se référer à Vose (2000, p. 82).

2.4. Conclusion du processus hypergeometrique

On utilisera ce processus dans notre domaine pour modéliser le tirage au sort d'animaux dans une population de faible taille (par exemple, un cheptel ou un petit lot d'aliments).

3. LE PROCESSUS DE POISSON

Expérience type : Des évènements particuliers sont observés : nouveaux cas d'une maladie, carcasses saisies sur une chaîne d'abatage, nombre de bactéries dans un volume de lait.... Le nombre moyen d'évènements particuliers par unité de temps (ou de volume) est λ. On s'intéresse au nombre d'évènements se produisant pendant un certain temps (ou dans un certain volume) t. Ce processus est dit « Poissonnien » s'il possède les cinq caractéristiques suivantes (si l'on est face à un exemple non proposé dans ce chapitre, il suffira de se demander si l'expérience que l'on désire simuler répond à ces cinq caractéristiques) :

- il modélise un nombre d'évènements x survenant dans un intervalle de temps (ou sur une surface, ou dans un volume) t ;
- il est caractérisé par un paramètre λ, le nombre moyen d'évènements par unité de temps (de surface, de volume, respectivement) ;
- la probabilité de survenue d'un évènement durant un intervalle de temps (de surface, de volume, respectivement) est **constante** durant tout le processus ;
- le nombre d'évènements qui survient durant un certain intervalle de temps (de surface, de volume, respectivement) est **indépendant du nombre d'évènements qui survient durant tout autre d'intervalle de temps** ;
- deux évènements ne peuvent survenir au même instant (ou au même endroit, respectivement)[5].

[5] Cette dernière condition peut être considérée comme accessoire dans notre domaine d'intérêt.

Encadré E : Loi inverse hypergéométrique

Ecriture *x ~ InvHypergéom(s, S, N)*

pour x entier $N - S \geq x \geq 0$

s : nombre de succès dans l'échantillon ($S \geq n \geq 0$)

S : nombre de succès dans la population ($N \geq S \geq 0$)

et N : taille de la population (> 0)

Densité

$$f(x) = \frac{C_S^{s-1} C_{N-S}^{x} (S - s + 1)}{C_N^{x+s-1} (N - x - s + 1)}$$

Fonction de répartition

$$F(x) = \sum_{y=0}^{x} \frac{C_S^{s-1} C_{N-S}^{y} (S - s + 1)}{C_N^{x+s-1} (N - y - s + 1)}$$

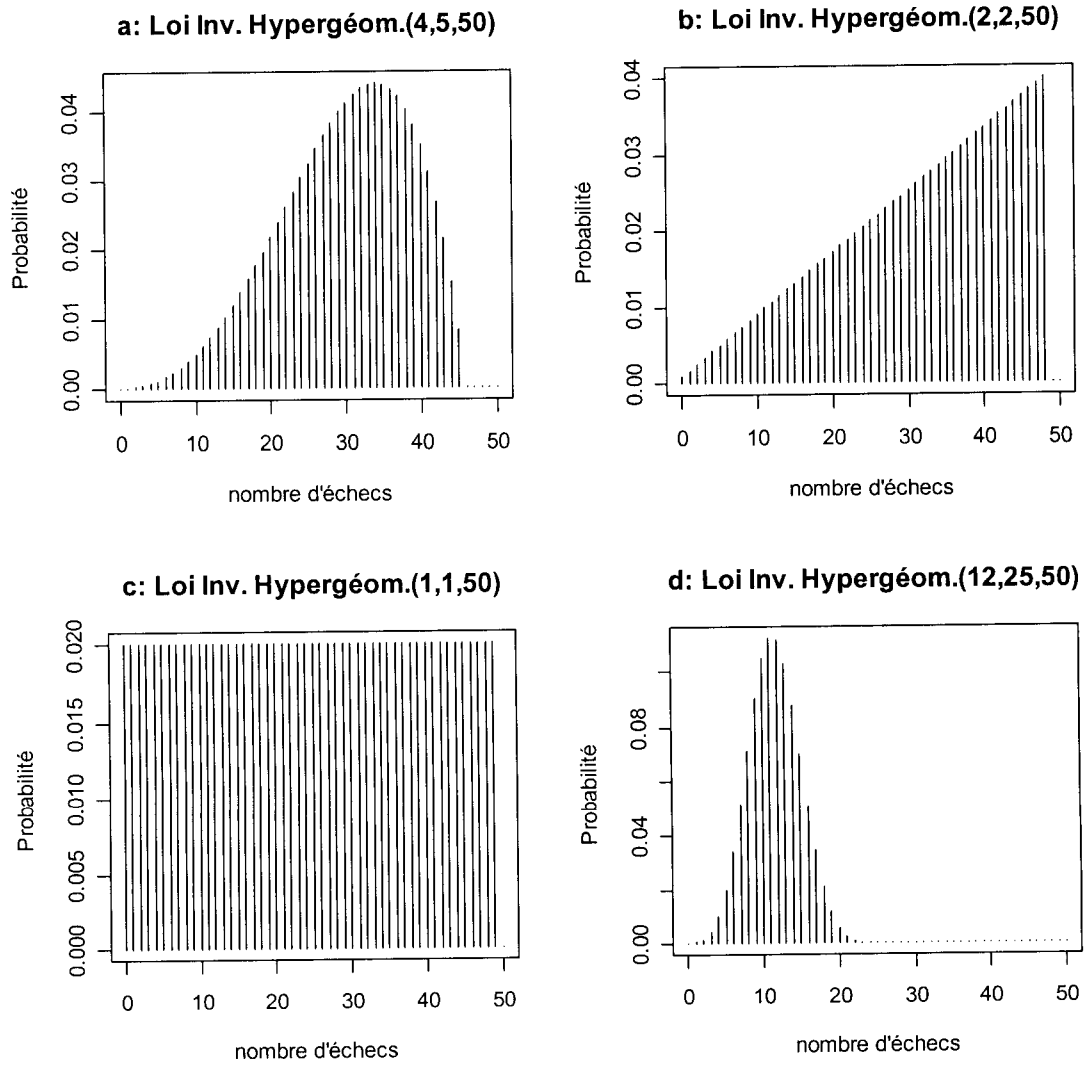
Propriétés

- Si $N >> s$, $x \sim BinNeg(s, S/N)$
- Si $N >> S$, $x \sim gamma(s, S/N)$

Ecriture Excel© : non implémentée

Ecriture @Risk©: non implémentée

Exemple



Ce processus est souvent considéré comme adéquat pour modéliser tout évènement rare. Cependant, il s'agit d'une généralisation qui n'a pas lieu d'être sans discussion : il est plus juste de considérer ce processus comme permettant de modéliser tout phénomène, même fréquent, dit « **sans mémoire** », *i.e.* **lorsque l'occurrence de chaque événement est indépendant de l'occurrence des autres évènements**. Ceci supprime dans notre domaine d'intérêt tout processus réagissant par agrégats, et donc toute maladie contagieuse, même rare. Par exemple, le processus de Poisson pourra être utilisé pour simuler le nombre de cas d'une maladie sporadique type charbon bactéridien ; il ne serait pas adéquat pour simuler la propagation d'une maladie du type fièvre aphteuse sur une zone. Il peut être utilisé pour simuler la distribution d'un nombre de bactéries dans un volume si celles-ci ne se présentent pas sous forme d'agrégats, mais sont réparties de façon homogène dans un prélèvement, quelle que soit la taille de celui-ci.

Un processus de Poisson est donc caractérisé par trois valeurs : x le nombre d'évènements, λ le nombre moyen d'évènements par unités de temps, d'espace ou de volume, et t le temps, l'espace ou le volume considéré. La connaissance de deux de ces paramètres permettra de faire des inférences sur le troisième (tableau III).

TABLEAU III

Relations entre les trois caractéristiques d'un processus de Poisson

Paramètre connus	Loi de la troisième caractéristique
λ, t	*Poisson*
λ, x	*gamma*
x, t	*gamma**

* : une des lois possibles (*cf.* infra)

3.1. NOMBRE D'EVENEMENTS OBSERVES : LA LOI DE POISSON

Si le nombre moyen d'événements par unité de temps, d'espace ou de volume est λ et que l'on désire modéliser le nombre d'évènements survenant pendant un intervalle de temps, ou une surface ou un volume t, on a :

$x \sim Poisson(\lambda t)$,

où $Poisson(\lambda t)$ est une loi de Poisson de paramètre λt (*cf.* encadré F). Les conditions d'application limitantes de cette loi sont 1) la probabilité d'occurrence d'un événement est constante lors de tout le processus ; 2) les évènements sont strictement indépendants.

Applications

1. On pourra appliquer la loi de Poisson pour simuler l'occurrence annuelle d'une maladie non (ou peu) transmissible.
2. On pourra appliquer la loi de Poisson pour simuler l'occurrence sur une surface donnée d'une maladie, non (ou peu) transmissible.

Nota : Si $X \sim Poisson(\lambda)$, l'intervalle de temps (d'espace, de volume) moyen entre deux évènements est $\beta = 1 / \lambda$.

Propriétés utiles

1. La somme de deux variables aléatoires de Poisson **indépendantes** est une variable aléatoire de Poisson. Si $X_1 \sim Poisson(\lambda_1)$ et $X_2 \sim Poisson(\lambda_2)$ alors $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

   Exemple : si le nombre annuel d'animaux nouvellement affectés par le tétanos dans une région suit une loi de Poisson de paramètre 5 et le nombre annuel d'animaux nouvellement affectés par le tétanos dans une autre région suit une loi de Poisson de paramètre 10, la somme du nombre annuel d'animaux nouvellement affectés par le tétanos dans ces deux régions suit une loi de Poisson de paramètre 15.

Encadré F : Loi de Poisson

Ecriture $x \sim Poisson(\lambda t)$

avec λ nombre moyen d'événements par unité de temps (d'espace ou de volume) considéré et t le temps (l'espace ou le volume) considéré.

Espérance λt

Variance λt

Densité $f(x) = \exp(-\lambda t)\frac{(\lambda t)^x}{x!}$

Fonction de répartition $F(x) = e^{-\lambda t}\sum_{i=0}^{x}\frac{(\lambda t)^i}{i!}$

Propriétés

- si $x_1 \sim Poisson(\Lambda_1)$, $x_2 \sim Poisson(\Lambda_2)$ et x_1 et x_2 indépendants, alors $x_1 + x_2 \sim Poisson(\Lambda_1 + \Lambda_2)$
- si $x \sim Binom(n, p)$, n est grand (> 40) et np petit (<5), alors $x \sim Poisson(np)$
- si $x \sim Poisson(\Lambda)$ et Λ grand (>5), alors $x \sim N(\Lambda, \sqrt{\Lambda})$

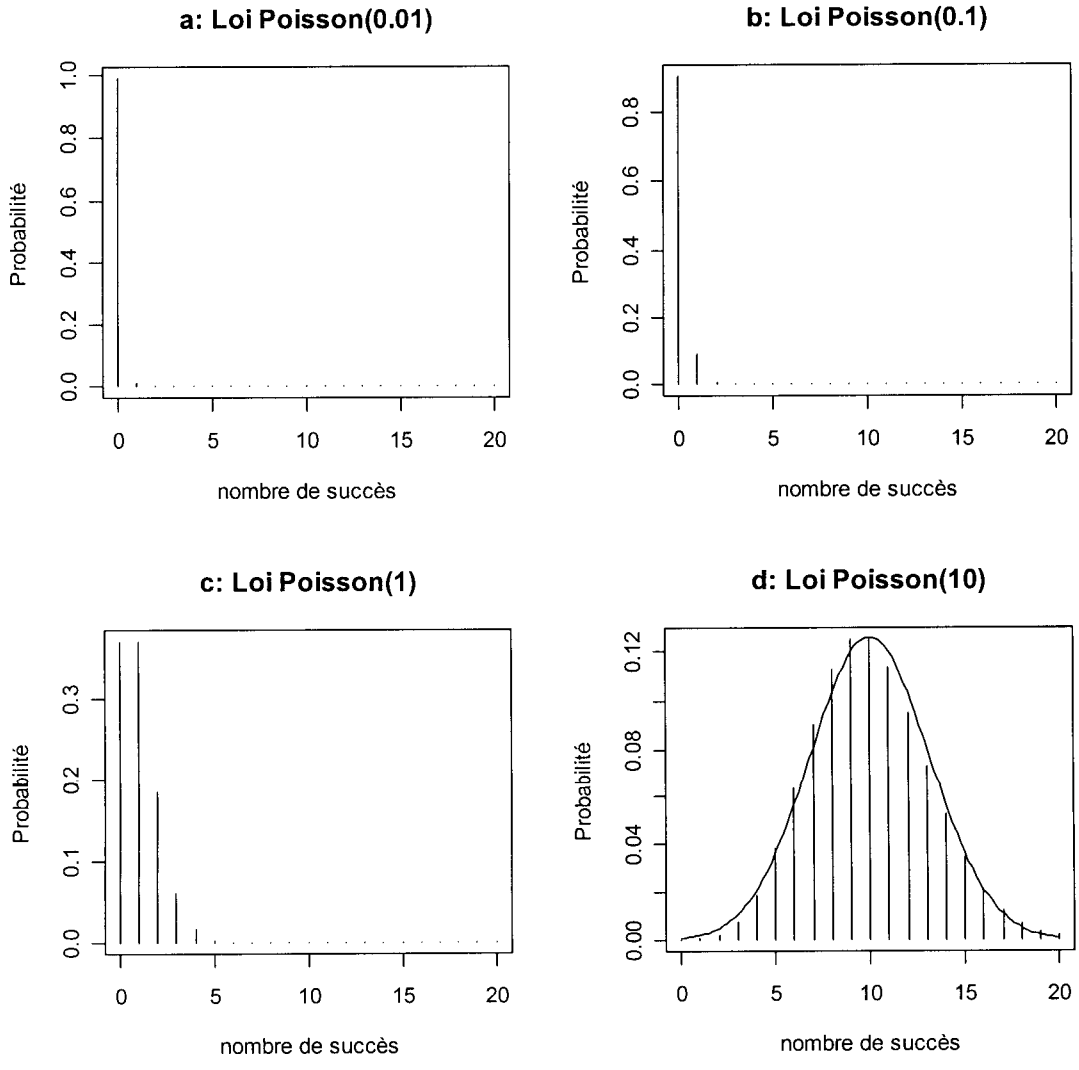
Ecriture Excel©

densité = LOI.POISSON(x, λt, 0)

fonction de répartition = LOI.POISSON(x, λt, 1)

Ecriture @Risk© = RISKPOISSON(λt)

Exemples : la figure d représente l'approximation de la loi de Poisson vers la loi normale.



2. Corollaire : si le nombre annuel de nouveaux cas d'une maladie non contagieuse suit une loi de Poisson de paramètre λ, le nombre de cas mensuel suit une loi de Poisson de paramètre $\lambda/12$, le nombre de cas quotidien suit une loi de Poisson de paramètre $\lambda/365$, etc.
3. La loi de Poisson est la loi vers laquelle converge la loi binomiale lorsque la probabilité p de succès est très faible, et lorsque la taille de l'échantillon n est grand (en pratique, on utilise comme limite d'utilisation $n > 40$, $np < 5$. **Exemple** : le nombre d'individus infectés par une maladie rare (p petit) dans un échantillon grand (n grand) suit une loi de Poisson de paramètre np (avec la limite d'indépendance entre les évènements).
4. La loi de Poisson tend vers la loi normale lorsque λt est grand (> 5).

3.2. TEMPS NECESSAIRE A L'OBSERVATION DE *X* EVENEMENTS : LOI GAMMA

Si le nombre moyen d'événements par unité de temps, d'espace ou de volume est λ et que l'on désire modéliser le temps nécessaire à l'observation des x prochains évènements, on a :

$t \sim gamma(x, 1/\lambda)$

Ou $gamma(a, b)$ est la loi gamma de paramètre de forme a et de paramètre d'échelle b (*cf.* encadré G).

Applications

1. On cherche à évaluer le temps nécessaire pour observer un certain nombre de foyers d'une maladie sporadique.
2. On cherche à évaluer la taille de la zone à étudier pour observer un certain nombre de foyers d'une maladie sporadique.

Exemple 1 : On peut penser que le nombre annuel de cas d'une maladie non infectieuse suit une loi de Poisson de paramètre $\lambda = 5$. Le temps avant l'observation **d'un** cas suit une loi gamma de paramètre $a = 1$, $b = 1/5$, l'inverse du paramètre de la loi de Poisson. *Nota* : toute loi gamma de paramètre de forme $a = 1$ est aussi connue sous le nom de « loi exponentielle ».

Exemple 2 : On peut penser que le nombre annuel de cas d'une maladie non infectieuse suit une loi de Poisson de paramètre $\lambda = 5$. Le temps avant l'observation de cinq cas suit une loi gamma de paramètre $a = 5$, le nombre de cas attendu et $b = 1/5$, l'inverse du paramètre de la loi de Poisson.

3.3. INCERTITUDE SUR LE NOMBRE MOYEN D'EVENEMENTS PAR UNITE DE TEMPS : LA LOI GAMMA

Si l'on a observé, durant un intervalle de temps t, x évènements, on peut estimer le nombre moyen d'évènements par unité de temps selon $\hat{\lambda} = x/t$. L'incertitude sur ce nombre moyen d'événements par unité de temps λ peut être modélisée (après application du théorème de Bayes[6], démonstration non proposée [*cf.* Vose, 2000] :

$\lambda \sim gamma(x, 1/t)$

Application

Si un paramètre suit une loi de Poisson, et que l'on dispose d'une estimation du paramètre de cette loi de Poisson, la loi gamma permet de modéliser l'incertitude de ce paramètre.

Exemple : On peut penser que le nombre annuel de cas d'une maladie suit une loi de Poisson. On a observé dans un pays, durant cinq années, 10 cas de cette maladie. L'incertitude sur le nombre moyen de cas annuel peut être modélisée selon une loi gamma de paramètres $a = 10$, le nombre de cas observé et $b = 1/5$, l'inverse du nombre d'années d'observation.

4. CONCLUSION

La connaissance des processus binomial, hypergéométrique et poissonnien permet de résoudre bon nombre de problèmes de caractérisation de paramètres en AQR liés à l'importation des animaux.

[6] en supposant une distribution *a priori* de λ uniforme.

Encadré G : Loi gamma

Ecriture *gamma*(*a*, *b*)

avec *a* : paramètre de forme (>0)

et *b* : paramètre d'échelle (>0)

Espérance ab

Variance ab^2

Densité $$f(x) = \frac{x^{a-1}\exp(-x/b)}{b^a\Gamma(a)}$$

avec Γ la **fonction** gamma (*cf.* Vose, 2000, p103)

Propriétés

- si $a = 1$, *gamma*(1, *b*) est une loi exponentielle de paramètre *b*
- si $x_1 \sim gamma(a, b_1)$, $x_2 \sim gamma(a, b_2)$ et x_1 et x_2 indépendants, alors $x_1x_2 \sim gamma(a, b_1+b_2)$

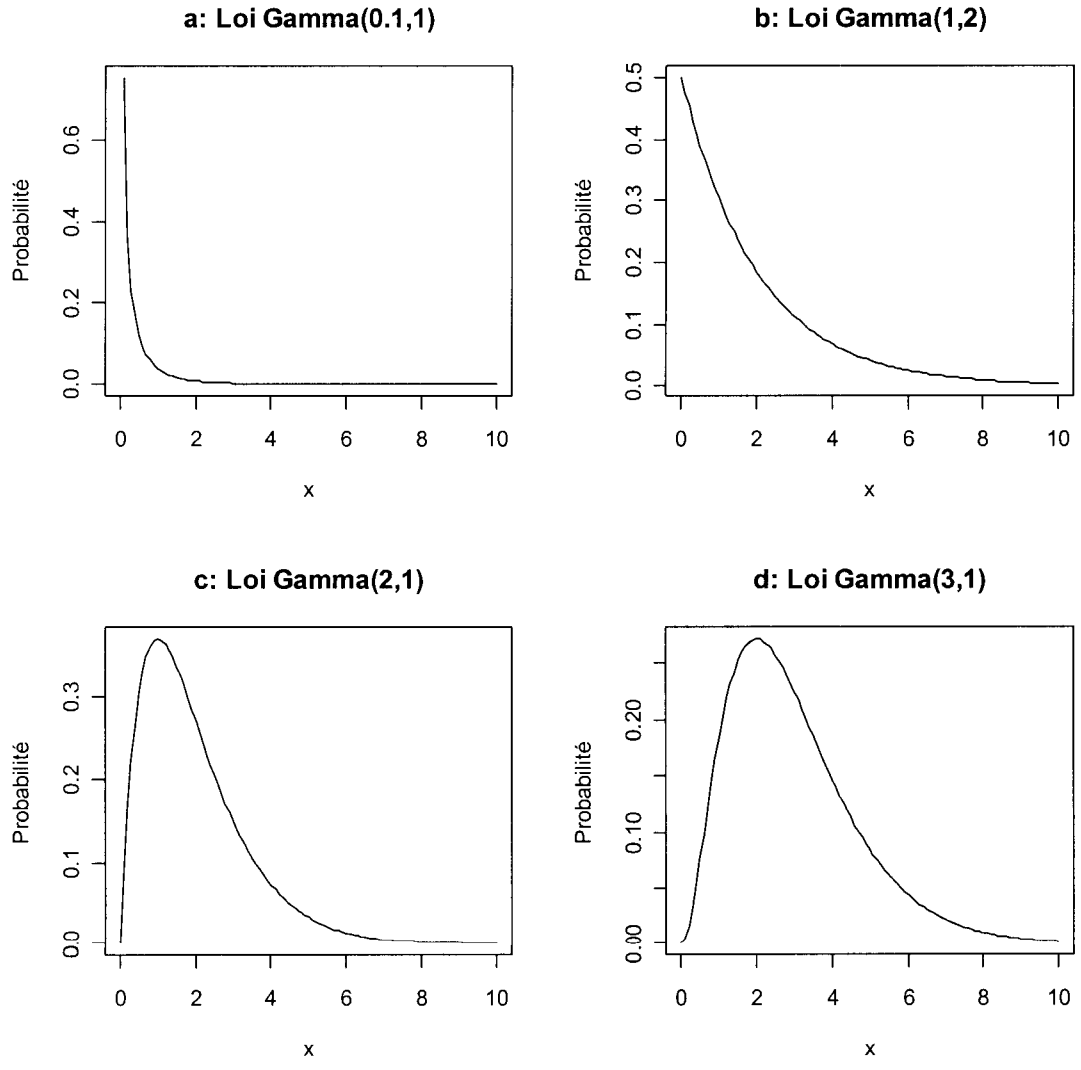
Ecriture Excel©

densité =LOI.GAMMA(*x*, *a*, *b*, 0)

fonction de répartition =LOI.GAMMA(*x*, *a*, *b*, 1)

Ecriture @Risk© =RISKGAMMA(*a*, *b*)

Exemples



II – DEUXIEME CAS
la théorie statistique permet de proposer une loi de distribution

Dans certains cas, la théorie statistique permet de déterminer la loi de distribution de paramètres à incorporer dans le modèle. C'est le cas notamment pour la définition de la distribution d'une moyenne d'une variable aléatoire ou de l'écart-type d'une variable aléatoire. Nous ne développerons ici que la modélisation de l'incertitude d'une moyenne.

Problème type : dans un modèle d'appréciation quantitative des risques, on utilise un paramètre μ qui représente la moyenne d'une variable aléatoire donnée. On ne dispose que d'un échantillon de n valeurs, réalisation de cette variable aléatoire, sur lequel on peut estimer la moyenne m et l'écart type s [*cf.* Pouillot et Sanaa, 2002]. La théorie statistique permet de proposer une modélisation de l'incertitude de l'estimation de μ, inhérente à la non exhaustivité de l'échantillon.

1. L'ECHANTILLON EST DE GRANDE TAILLE

On utilise généralement le seuil $n > 30$ (mais ce seuil n'est pas absolu). Dans ce cas, la théorie statistique nous permet d'utiliser la relation :

$$\mu \sim N\left(m, \frac{s}{\sqrt{n}}\right)$$

où $N(a, b)$ est la loi normale de paramètre a et b (*cf.* l'encadré H pour les généralités sur la loi normale, l'encadré I pour les caractéristiques de cette loi). On notera que cette relation est valable quelle que soit la distribution de la variable[7] dans la population[8].

Exemple : dans un modèle, un des paramètres nécessaires (que l'on nommera x) est la prévalence **moyenne** intra-cheptel de la maladie étudiée. On a mesuré les prévalences intra-cheptels dans 40 élevages infectés représentatifs de la population des élevages infectés. L'histogramme des valeurs observées est présenté figure 1a). La moyenne des prévalences observées est de 6,0%, l'écart-type des prévalences observées est de 2,4%. 6,0% est l'estimation la plus probable de x. Cependant, n'ayant réalisé cette estimation que sur un nombre limité de valeurs, il existe une certaine incertitude sur l'estimation de ce paramètre. La relation précédente nous permet de modéliser l'incertitude sur le paramètre x selon :

$$x \sim N\left(0{,}06, \frac{0{,}024}{\sqrt{40}}\right)$$

La distribution utilisée est proposée figure 1b). On notera que la distribution des valeurs observées ne semblait pas de type « normale » : comme n est « grand » (40), ceci n'a que peu d'importance.

[7] Ne pas confondre distribution de la **variable** (qui peut être quelconque) et distribution de la **moyenne de la variable** qui tend vers la loi normale.

[8] Cependant, le seuil de 30 observations est influencé par la distribution de la variable. Pour simplifier, plus la distribution de la variable « ressemble » à la loi normale, moins il faudra d'observations. Pour certaines lois très dissymétriques (loi gamma, loi binomiale négative,...) le seuil de 30 est largement insuffisant.

Encadré H : Généralités sur la loi normale (ou loi de Laplace-Gauss)

La loi normale est la loi de distribution la plus connue. Elle est d'une importance majeure en théorie statistique. Son utilisation en AQR liés à l'importation est plus limitée.

Applications

1. Le théorème central limite montre que la **somme** de n observations indépendantes **provenant d'une distribution quelconque** de moyenne μ et de variance σ^2 tend vers une loi normale de moyenne $n\mu$ et de variance $n\sigma^2$.
2. Par conséquent, la **moyenne** de n observations provenant d'une distribution quelconque de moyenne μ et de variance σ^2 tend vers une loi normale de moyenne μ et de variance σ^2/n.
3. Il est fréquent que certaines variables dans une population suivent une loi normale : taille des individus, poids des individus, etc.
4. Il s'agit de la loi vers laquelle converge la loi binomiale (lorsque np est grand)et la loi de Poisson (lorsque λ est grand).

Propriété utile

1. La somme de deux lois normales indépendantes est une loi normale : si $X \sim N(\mu_1, \sigma_1)$ et $Y \sim N(\mu_2, \sigma_2)$, alors $X + Y \sim N(\mu_1+\mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$. **Exemple** : si le nombre de bovins dans les cheptels peut être modélisé selon une loi normale de paramètre $\mu_{bov} = 40$ et $\sigma_{bov} = 10$ et le nombre d'ovins dans les cheptels peut être modélisé selon une loi normale de paramètre $\mu_{ov} = 100$ et $\sigma_{ov} = 30$, et si l'on considère que le nombre de bovins est indépendant du nombre d'ovins dans un cheptel (cette hypothèse est osée), le nombre d'animaux dans le cheptel suit une loi normale de paramètre $\mu = 140$, $\sigma = 31{,}6$.
2. La différence de deux loi normales indépendantes est une loi normale : si $X \sim N(\mu_1, \sigma_1)$ et $Y \sim N(\mu_2, \sigma_2)$, alors $X - Y \sim N(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$

Remarques :

- La loi normale peut en théorie simuler des valeurs comprises entre $-\infty$ et $+\infty$. Or, il est rare qu'un paramètre puisse varier sur une telle étendue (ex : taille ou poids des individus). En pratique, la loi normale est assez centrée : 99,9% des valeurs sont comprises dans un intervalle moyenne ± 3 écart-types, ce qui limite cet inconvénient. Il est cependant possible d'utiliser une loi « normale tronquée » définie sur un intervalle de valeurs plus raisonnable que la loi normale. Attention : les statistiques de la loi normale tronquée (moyenne, écart-type, médiane, percentiles, etc.) sont différentes de celles de la loi normale.
- La loi normale est une loi continue : or, elle peut être utilisée pour simuler la distribution d'un paramètre entier (ex : taille de cheptel). Il est possible d'utiliser l'arrondi d'une loi normale pour simuler ce type de données. Attention : les statistiques de cette loi normale arrondie (moyenne, écart-type, médiane, percentiles, etc.) sont différentes de celles de la loi normale.

Encadré I : Loi normale

Ecriture $N(\mu, \sigma)$

avec μ : moyenne

et σ : écart-type

Espérance μ

Variance σ^2

Densité $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Propriétés

- si $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ et X et Y indépendants alors $X + Y \sim N(\mu_1+\mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
- si $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$ et X et Y indépendants alors $X - Y \sim N(\mu_1-\mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Ecriture Excel©

densité = LOI.NORMALE(x, m, s, 0)

fonction de répartition = LOI.NORMALE(x, m, s, 1)

Si $\Pr(X \leq x) = z$, alors x =LOI.NORMALE.INVERSE(z, m, s)

Ecriture @Risk© =RISKNORMAL(m, s)

Exemple

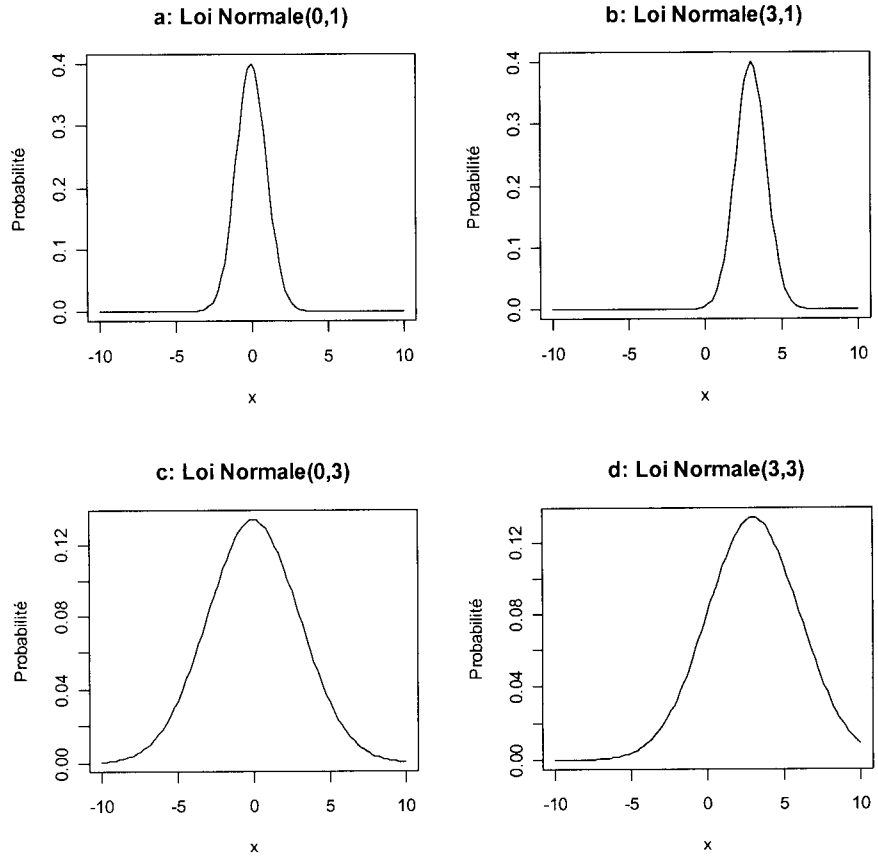
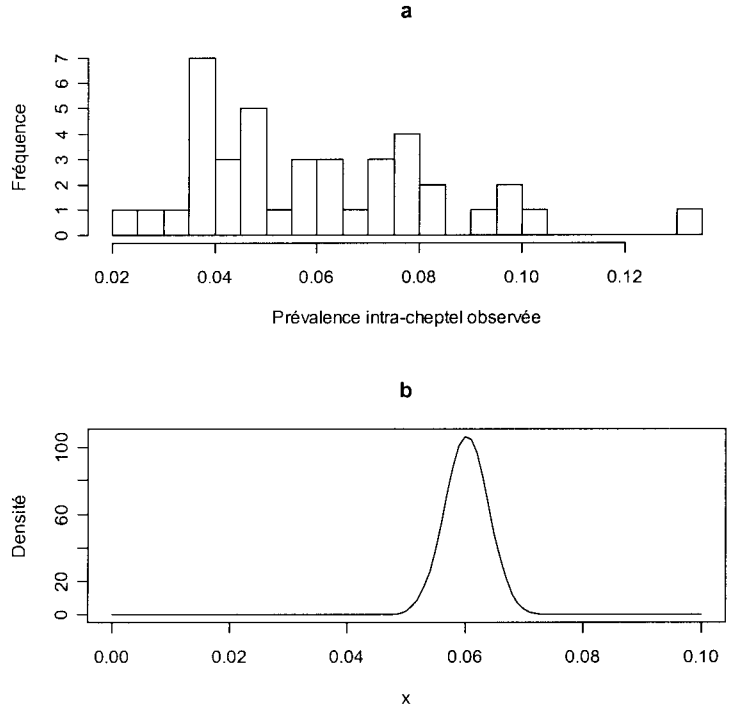


FIGURE 1

a) Histogramme des valeurs de prévalence intra-cheptel observées
b) Modélisation de la *moyenne* de la prévalence intra-cheptel



2. L'ECHANTILLON EST DE PETIT TAILLE

Dans ce cas, une autre condition doit être remplie : la distribution des **données** doit être normale. La théorie statistique nous permet, sous cette condition, d'utiliser la relation :

$$\mu \sim m + \left(\frac{s}{\sqrt{n}}\right) \times t_{n-1}$$

où t_{n-1} est la distribution t de Student à (n - 1) degrés de liberté[9].

Exemple : dans un modèle, un des paramètres nécessaires est la sensibilité individuelle moyenne d'un test, en conditions de terrain. On envoie une sérothèque de 100 sérums représentatifs de la population des animaux infectés à cinq laboratoires. Les sensibilités obtenues par ces laboratoires sont 0,87, 0,82, 0,82, 0,81 et 0,75 (soit une moyenne de 0,814 et un écart-type de 0,042).

Selon la relation précédente, et en supposant (ce qui est démontrable[10]) que la sensibilité observée par les laboratoires suit une loi de type normal, on peut utiliser la relation :

$$\mu \sim 0{,}814 + \frac{0{,}042}{\sqrt{5}} \times t_4$$

La figure 2 représente la distribution correspondante.

---

9 Nous ne présenterons pas cette distribution, en raison de son utilisation assez peu fréquente dans le domaine. Le lecteur intéressé est invité à se référer à tout ouvrage de statistique de base.

10 Le raisonnement pourrait être le suivant : nous sommes face à un processus binomial (dans la population de n = 100 sérums infectés, une certaine proportion *Se* réagissent au test, et on observe r réagissants). On a donc :

$$r \sim Binom(100, Se)$$

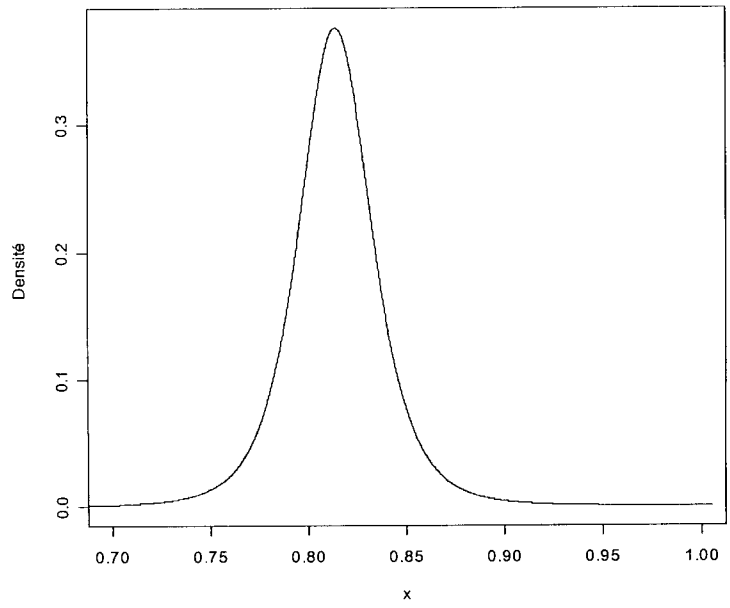
En raison de la convergence de la loi binomiale vers la loi normale pour $n \times Se$ grand (ce qui est le cas ici : $n \times Se \approx 80$), on a approximativement :

$$R \sim N(100\,Se, \sqrt{100\,Se(1-Se)})$$

La sensibilité est estimée par $x = r/100$. Donc l'estimation de la sensibilité suit une loi normale.

FIGURE 2

Modélisation de l'incertitude sur la sensibilité du test (cf. texte)



3. CONCLUSION

Nous n'avons présenté ici qu'un nombre très limité, à titre d'exemple, d'utilisations de la théorie statistique pour la caractérisation de distribution de variables aléatoires entrant dans un modèle d'appréciation quantitative des risques. Il est également possible d'utiliser des relations statistiques classiques pour la caractérisation de l'incertitude liée à l'estimation d'un écart-type, d'un pourcentage, etc.

Il est possible de faire référence à des théories statistiques complémentaires d'un grand intérêt dans le domaine de l'appréciation quantitative des risques : la statistique Bayésienne, ou la théorie du Bootstrap, particulièrement intéressantes pour la spécification de distributions correspondant à l'incertitude d'un paramètre. Ces développements sortent cependant du cadre de cet article de généralités.

III - TROISIEME CAS

on dispose d'un grand nombre de données observées

Dans certains cas, on dispose d'un grand nombre de données observées. La définition de la loi de distribution de la variable est alors possible selon deux modalités principales : la loi empirique ou l'ajustement d'une loi paramétrique.

1. LOI EMPIRIQUE

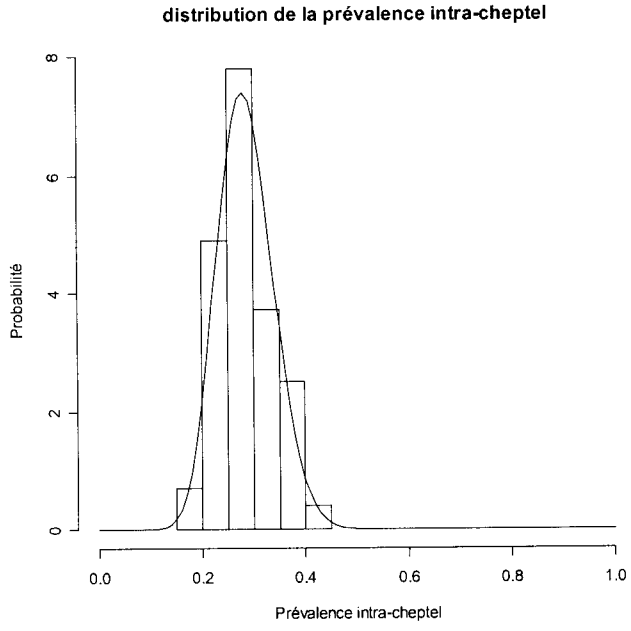
Si l'on dispose d'un grand nombre d'observations, il est possible, afin de simuler des réalisations de la variable aléatoire, de ne pas utiliser une loi de distribution « paramétrique » (c'est-à-dire définie mathématiquement à l'aide d'un nombre limité de paramètres), mais de tirer au sort directement les valeurs nécessaires parmi les valeurs observées. Cette méthode a pour avantages d'être simple et de ne pas entraîner de biais, possibles lors d'un ajustement non adéquat à une loi de distribution particulière. Elle a pour inconvénient de ne pas refléter l'ensemble des possibilités, mais un échantillon : il ne sera par exemple pas possible de prendre en compte une possibilité

plus extrême que la possibilité la plus extrême **observée**.

Exemple : on a mesuré les prévalences intra-cheptel dans 200 cheptels infectés : l'histogramme des valeurs obtenues est proposé figure 3. Pour simuler des prévalences intra-cheptel, on tirera au sort des valeurs parmi les valeurs observées.

FIGURE 3

Exemple d'utilisation d'un grand nombre de données observées : l'histogramme présente la distribution des données observées : cette loi de probabilité peut être utilisée directement dans l'analyse. Il est également possible d'ajuster sur ces données une loi de probabilité connue, telle que celle représentée par la courbe (ici, une loi bêta).



2. AJUSTEMENT DES DONNEES PAR UNE LOI DE DISTRIBUTION

Si l'on dispose d'un grand nombre de données observées, il est possible de chercher la loi de distribution de probabilité paramétrique (*i.e.* définie mathématiquement telle que la loi normale, la loi log normale ou la loi gamma, etc.) qui s'ajuste le mieux à ces données. Cette procédure est longue et complexe. Cependant, les deux principaux logiciels utilisés en analyse de risque (@Risk et Crystal Ball) proposent ces procédures d'ajustement. La méthode est la suivante :

1. on choisit *a priori* un certain nombre de distributions paramétriques susceptibles de s'ajuster aux données observées. Ce choix est principalement lié :

   a. au caractère continu ou discret du paramètre :

      - continu : lois normale, log-normale, gamma, bêta, etc.
      - discret : loi de Poisson, loi binomiale, loi binomiale négative, etc.

   b. au domaine de définition : la loi bêta ne peut modéliser que des variables comprises dans l'intervalle [0 ; 1], la loi gamma que dans l'intervalle [0, + ∞[, etc.

   c. parfois, aux connaissances (bibliographiques, par exemple) de la loi suivie par le type de variable considéré : la variabilité de la concentration en un grand nombre de contaminants chimiques dans des aliments est ainsi souvent représentée par une loi log-normale.

   Cette étape est facultative : on peut décider d'ajuster toutes les lois proposées par le logiciel. Cependant, une vérification en fin de procédure de la plausibilité de la loi proposée par le logiciel sera nécessaire.

2. pour chacune de ces distributions, on choisit les paramètres qui permettront à la distribution de s'ajuster au mieux aux données. Deux procédures statistiques classiques sont utilisées :

a. la méthode des moments : elle consiste à choisir comme paramètres de la distribution recherchée les valeurs permettant d'ajuster la moyenne des données observées à la moyenne de la distribution recherchée et l'écart-type (ou son carré : la variance) des données observées à l'écart-type (ou la variance) de la distribution recherchée. **Exemple** : on dispose d'une série de données. On calcule la moyenne m et l'écart-type s de ces données. On veut ajuster une loi gamma de paramètres a et b à ces données : on sait que la moyenne d'une loi gamma(a, b) est ab, et que son écart-type est $b\sqrt{a}$. On choisira a et b tels que $m = ab$ et $s = b\sqrt{a}$, c'est-à-dire $a = (m / s)^2$ et $b = s^2/m$. Cette méthode des moments est la plus simple des méthodes d'ajustement, mais elle se révèle souvent peu performante ;

b. la méthode de maximum de vraisemblance : selon les paramètres de la loi de distribution recherchée, chaque valeur observée à une certaine probabilité d'être obtenue. Par exemple, une valeur de 10 aurait une très faible probabilité d'être observée si le paramètre suivait une loi normale de paramètre $m = 0$, $s = 1$, mais une plus grande probabilité si le paramètre suivait une loi normale de paramètre $m = 9$, $s = 2$. La méthode de maximum de vraisemblance consiste à rechercher les paramètres qui permettraient d'obtenir la plus forte probabilité d'observer l'ensemble des données observées ;

3. on cherche ensuite, parmi l'ensemble des distributions paramétrées celle qui s'ajuste le mieux aux données, à l'aide d'une statistique permettant de juger de cette adéquation. Ces tests sont principalement le test du Chi deux, le test de Kolmogorov-Smirnov, et le test d'Anderson-Darling. Le test d'Anderson Darling semble le plus approprié.

Exemple : à partir des données de prévalence intra-cheptel observées (histogramme proposé figure 3), on a ajusté un ensemble de lois paramétriques. Le meilleur ajustement est obtenu par une loi bêta (*cf.* figure 3).

Nota : L'ajustement de lois paramétriques pour des lois de contamination est souvent optimal avec une loi log-normale. Nous proposons dans l'encadré J les caractéristiques de cette loi.

3. CONCLUSION

Quand on dispose d'un grand nombre de données, la caractérisation d'une loi de distribution est confortable, car basée sur des réalités de terrain. L'ajustement par une loi paramétrique ne doit cependant pas être réalisé de façon automatisée sans jugement de l'adéquation de la forme de la loi au type de variable aléatoire.

Encadré J : Loi log-normale

Ecriture $X \sim LN(\mu, \sigma)$ si $\ln(X) \sim N(\mu, \sigma)$

pour $x > 0$

avec μ : moyenne de la loi **normale**

et σ : écart-type de la loi **normale** (>0)[11]

Espérance $\exp(\mu + \sigma^2 / 2)$

Variance $\sqrt{\exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)}$

Densité $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$$

Estimation des paramètres (méthodes des moments)

$\hat{\mu}$: moyenne de ln(***X***)

$\hat{\sigma}$: écart-type de ln(***X***)

avec ***X*** un jeu de données observées.

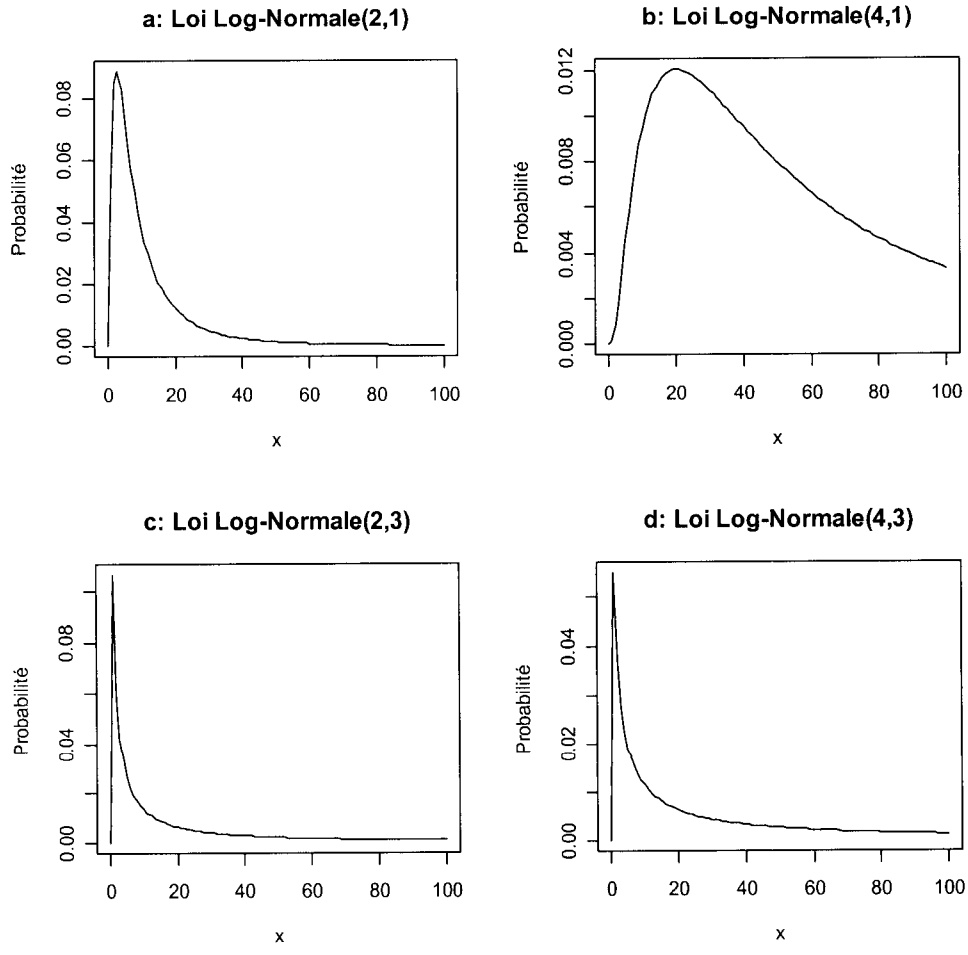
Ecriture Excel©

densité = LOI.LOGNORMALE(x, μ, σ)

Si $\Pr(X \leq x) = z$, alors x =LOI.LOGNORMALE.INVERSE(z, μ, σ)

Ecriture @Risk© = RISKLOGNORM(m, s)[12]

Exemple



[11] Il s'agit d'un type de paramétrisation possible. Le second est $X \sim LN(m, s)$ avec m moyenne de la loi log-normale, et s écart-type de la loi log-normale.

[12] Attention : paramétrisation telle que définie dans la précédente note de bas de page.

IV - QUATRIEME CAS

on ne dispose pas de données : recours à des dires d'experts

Dans ce cas, il est nécessaire de définir la loi de distribution et ses paramètres à partir de dires d'experts. Deux cas peuvent se présenter : soit on connaît la forme de la loi de distribution de la variable aléatoire, et on ne cherche que ses paramètres, soit on ne sait rien de la forme ni des paramètres. On ne présentera pas ici les procédures permettant d'extraire, au mieux, des connaissances d'expert. Des méthodes particulières, type méthode DELPHI, ont été développées dans ce sens.

1. ON CONNAIT LA FORME DE LA LOI DE DISTRIBUTION

Si le processus d'obtention du paramètre est connu, on pourra utiliser la loi de distribution correspondante, et définir ses paramètres à partir de dires d'experts. Cette procédure est plus ou moins facile, selon le paramètre et la distribution choisie. On s'attachera à demander aux experts des paramètres ayant un sens pratique.

Exemple : si l'on cherche les paramètres d'une loi normale, il est possible de demander aux experts la moyenne du paramètre, car il s'agit d'un paramètre bien appréhendé. En revanche, on ne demandera pas l'écart type, ou encore moins la variance car ce paramètre n'a pas de sens concret. On pourra en revanche demander l'intervalle de valeurs autour de la moyenne qui, selon eux, contient 95% des individus. On sait que, si $X \sim N(m, s)$, alors 95% des observations sont situées dans l'intervalle $m \pm 1{,}96\ s$: à partir de cet intervalle, on pourra alors obtenir s.

2. ON NE CONNAIT PAS LA FORME DE LA LOI DE DISTRIBUTION

Si le processus d'obtention du paramètre n'est pas connu, on pourra utiliser une des distributions particulières employées en analyse de risque (tableau IV). Ces distributions permettent de donner une forme à la densité de probabilité, en fonction d'un nombre limité de paramètres facilement appréhendés par les experts (minimum, maximum, valeur la plus probable, etc.). Elles sont en général très sensibles aux estimations des valeurs maximales et minimales. Elles ne sont pas « naturelles » car elles ne reflètent pas des distributions habituellement observées pour des phénomènes biologiques.

TABLEAU IV

Quelques lois de distribution utilisées pour la caractérisation de paramètres à partir de dires d'experts

Situation	Distribution possible
Les experts peuvent proposer des valeurs maximale et minimale et affirmer que toute valeur située dans cet intervalle semble **également probable**	*Unif*(minimum, maximum)
Les experts peuvent proposer des valeurs maximale et minimale et une **valeur la plus probable**	*Triang*(minimum, plus probable, maximum)
ou*	*Pert*(minimum, plus probable, maximum)
Les experts peuvent proposer les valeurs maximale et minimale et les probabilités pour un **certain nombre de valeurs**	*Générale*(minimum, maximum, vecteur de valeurs, vecteur de probabilités)

* La distribution de Pert est moins sensible aux valeurs extrêmes et a une forme plus naturelle.

3. LOI UNIFORME

La loi uniforme est utilisée pour modéliser un paramètre en évaluation des risques lorsque l'on connaît la valeur minimale et la valeur maximale possible de la variable et que, selon les experts, **toute valeur dans cet intervalle est également probable**. La connaissance de deux paramètres est nécessaire : le minium et le maximum. On écrit :

$X \sim Unif(a, b)$

avec a la valeur minimale et b la valeur maximale (*cf* encadré K).

4. LOI TRIANGULAIRE

La loi triangulaire est utilisée en évaluation des risques lorsqu'il est possible de définir une valeur minimale, une valeur maximale et une valeur la plus probable pour le paramètre. Comme son nom l'indique, la densité de cette distribution a une forme de triangle.

$X \sim Triang(a, b, c)$

avec a la valeur minimale, b la valeur la plus probable (mode de la distribution) et c la valeur maximale (*cf.* encadré L).

5. LOI PERT

La loi Pert est utilisée en évaluation des risques comme la loi triangulaire, lorsqu'il est possible de définir une valeur minimale, une valeur maximale et une valeur la plus probable pour le paramètre. La densité de cette distribution a une forme plus « arrondie » : elle est plus plausible que la loi triangulaire. Cette loi est dérivée de la loi bêta présentée précédemment, elle est paramétrée de façon à ce que les valeurs à caractériser soient aisément interprétables par les experts. De plus, cette loi est beaucoup moins sensible que la loi triangulaire aux paramètres de minimum et de maximum. Elle s'écrit :

$X \sim Pert(a, b, c)$

avec a la valeur minimale, b la valeur la plus probable et c la valeur maximale (*cf.* encadré M).

6. LOI GENERALE

La loi générale est une loi très flexible, permettant aux experts de définir non seulement les valeurs minimales et maximales que peut prendre le paramètre, mais également des valeurs spécifiques de densités. Cette loi est proposée par le logiciel @risk.

Exemple : supposons qu'il soit demandé aux experts la loi de distribution de la taille des cheptels dans une zone. Les experts peuvent fournir la taille minimale (exemple : 1), la taille maximale (exemple : 140), et peuvent appréhender certaines valeurs de « densité »[13] pour certaines tailles de cheptel : par exemple, il peuvent penser que l'effectif le plus fréquent est 40, et attribuer une valeur de « densité » de 10 pour cette valeur. L'effectif 20 peut être proposé avec une densité de 7, l'effectif 70 à 4, l'effectif 100 à 2. La courbe de densité correspondante est présentée figure 4.

En conclusion, la caractérisation de lois de distribution à partir de dires d'experts nécessite de définir une forme de loi qui semble cohérente pour l'évaluateur. Il existe un grand nombre d'autres lois utilisées dans ce contexte (loi Trigen, Loi Pert modifiée, etc.). Le lecteur intéressé est invité à se référer aux manuels d'utilisation des logiciels dédiés à l'appréciation quantitative des risques.

[13] Le terme densité est proposé entre guillemets, car il ne correspond pas *sensu stricto* a une valeur de densité. L'algorithme du logiciel fera une paramétrisation correcte de ces valeurs, pour que l'aire sous la courbe de la loi de densité soit bien égale à 1.

Encadré K : Loi uniforme

Ecriture $X \sim Unif(a, b)$

pour $x \in [a, b]$

avec a : minimum, b : maximum

Espérance $\frac{a+b}{2}$

Variance $\frac{(a-b)^2}{12}$

Densité $f(x) = \frac{1}{b-a}$

Fonction de répartition $F(x) = \frac{x-a}{b-a}$

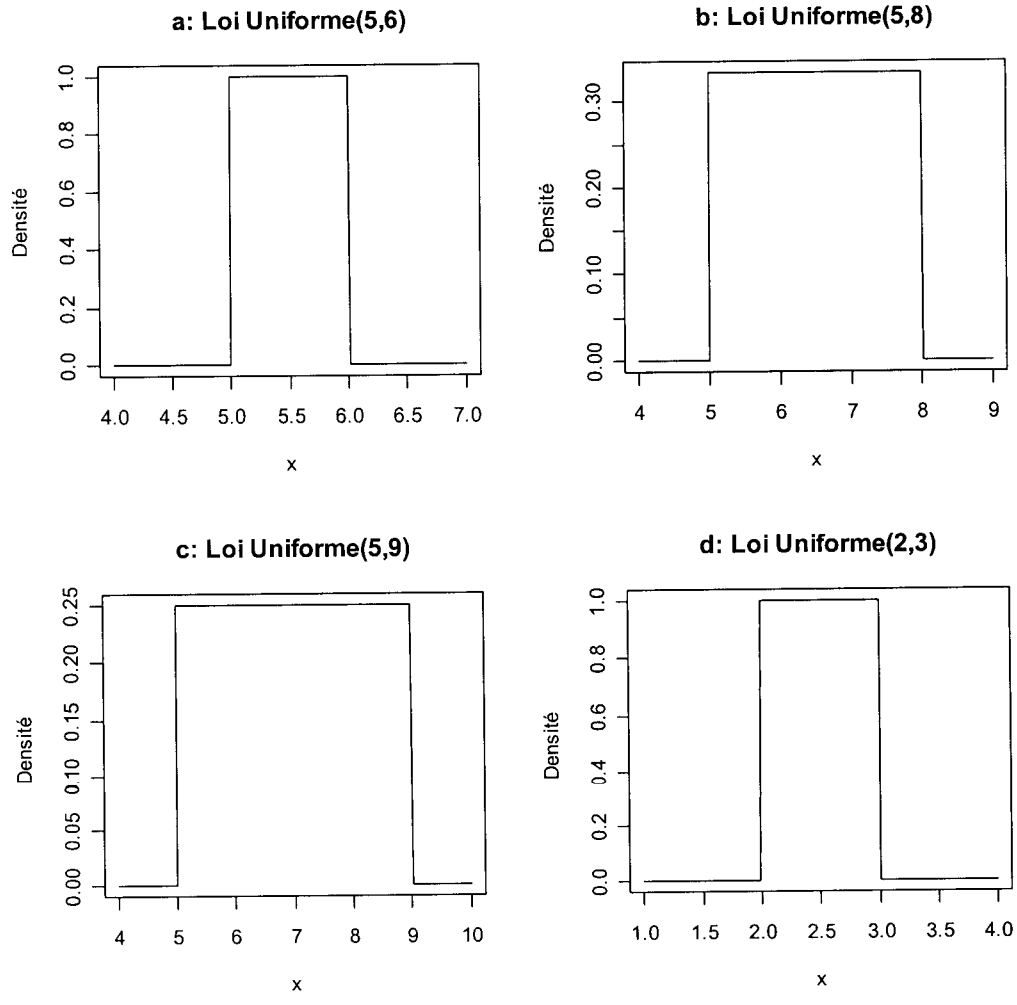
Ecriture Excel©

densité $= \frac{1}{b-a}$

fonction de répartition $= \frac{(x-a)}{(b-a)}$

Ecriture @Risk© = RISKUNIFORM(a, b)

Exemple



Encadré L : Loi triangulaire

Ecriture *X ~ Triang(a, b,c)*

pour $x \in [a, c]$

avec a : minimum, b : valeur la plus probable, et c maximum

Espérance

$$\frac{a+b+c}{3}$$

Variance

$$\frac{a^2+b^2+c^2-ab-ac-bc}{18}$$

Densité

$$f(x)=\frac{2(x-a)}{(b-a)(c-a)} \text{ si } a \le x \le b$$

$$f(x)=\frac{2(c-x)}{(c-a)(c-b)} \text{ si } b \le x \le c$$

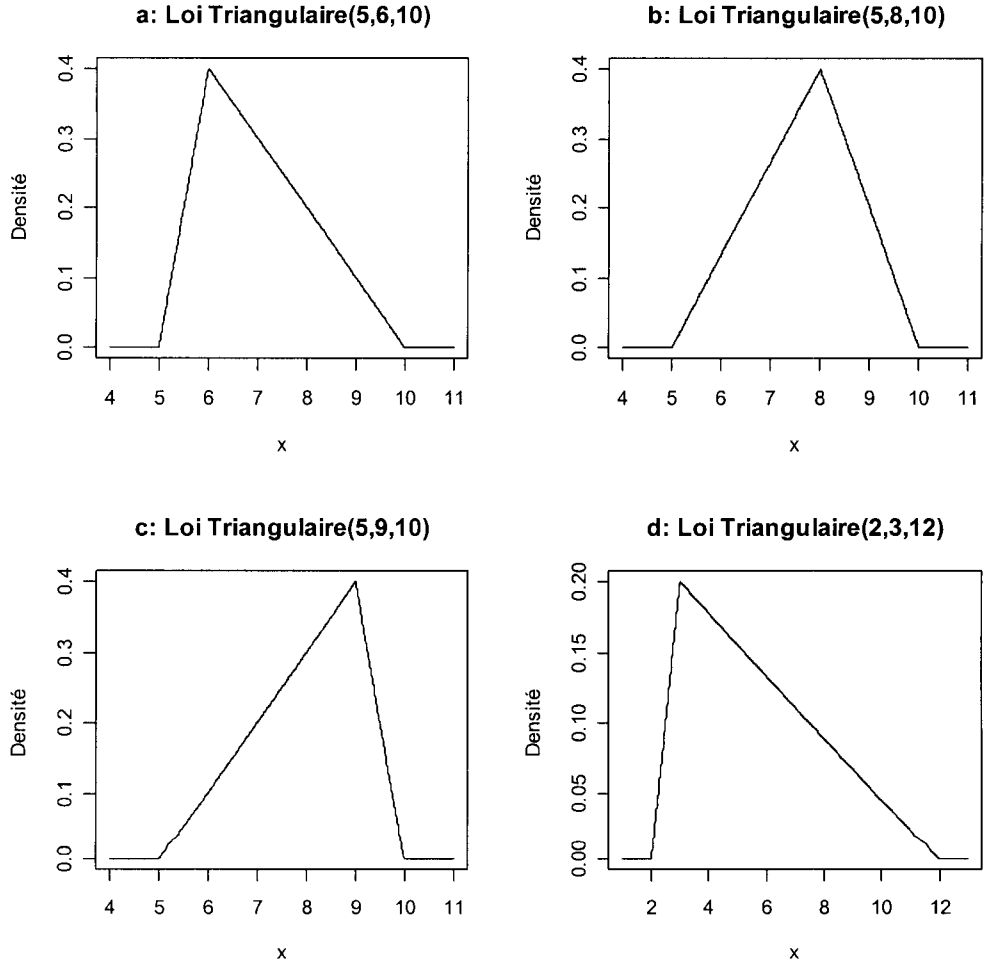
Fonction de répartition

$$F(x)=\frac{(x-a)^2}{(b-a)(c-a)} \text{ si } a \le x \le b$$

$$F(x)=1-\frac{(c-x)^2}{(c-a)(c-b)} \text{ si } b \le x \le c$$

Ecriture @Risk© = RISKTRIANG(*a*, *b*, *c*)

Exemples



Encadré M : Loi Pert

Ecriture *X ~ Pert(a, b, c)*

pour $x \in [a, c]$

avec *a* : minimum, *b* : valeur la plus probable, et *c* maximum

Espérance $\mu = \dfrac{a+4b+c}{6}$

Densité : la loi *Pert* est une dérivée de la loi *bêta*

si *X* ~Pert(*a*, *b*, *c*), alors *X* ~ *Beta*(α, β) × (*c* - *a*) + *a*, avec

$$\alpha = \frac{(\mu - a)(2b - a - c)}{(b - \mu)(c - a)}$$

$$\beta = \frac{\alpha(c - \mu)}{(\mu - a)}$$

μ étant la moyenne de la distribution *Pert*

Ecriture @Risk© =RISKPERT(*a*, *b*, *c*)

Exemple

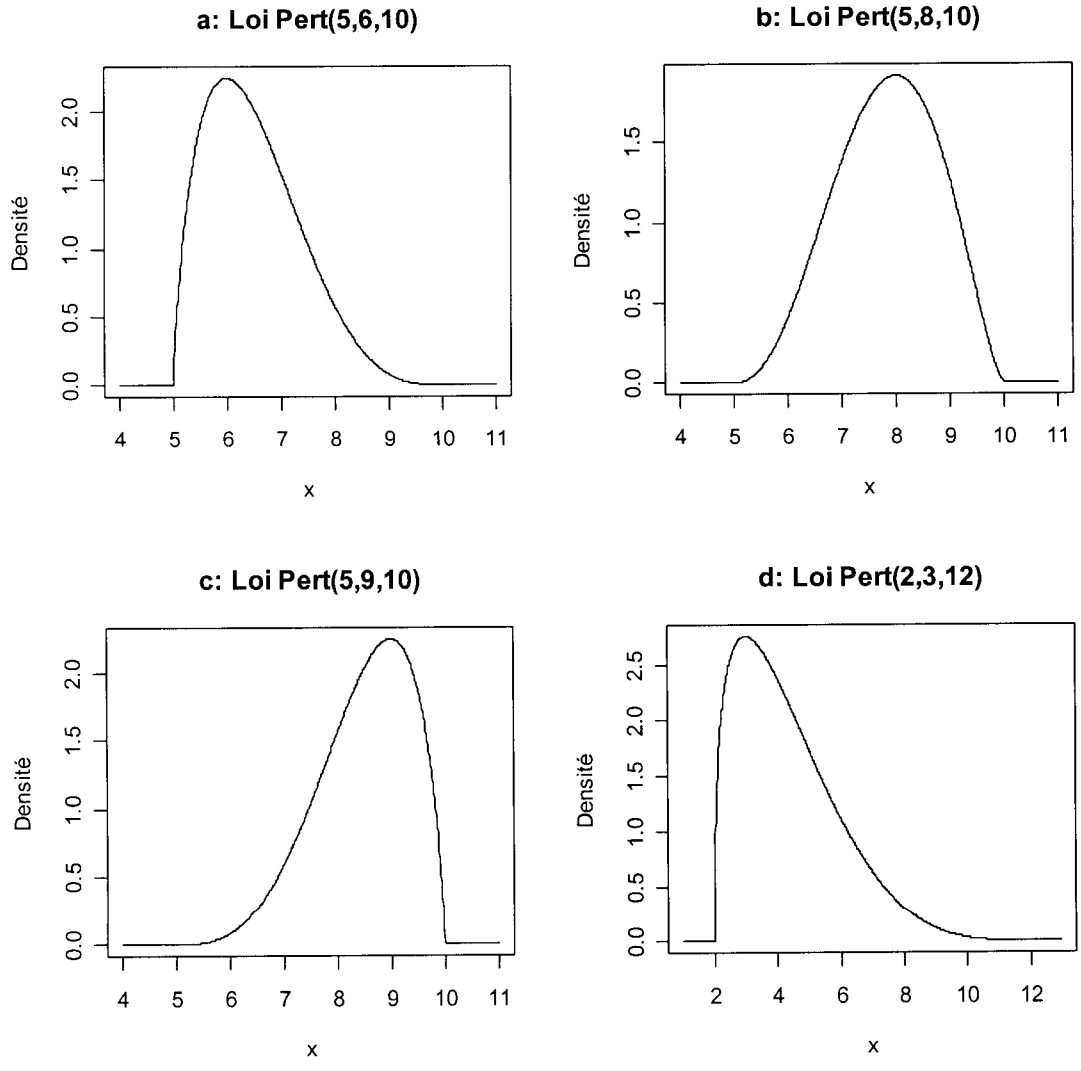
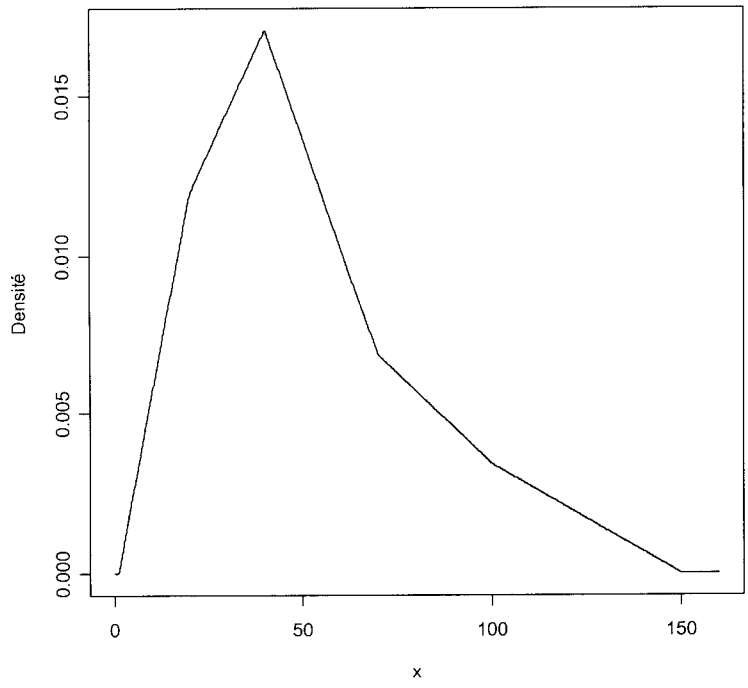


FIGURE 4

Loi générale(0, 150, {20, 40, 70, 100}, {7, 10, 4, 2})



V - CONCLUSION

La caractérisation des lois de densité dans un modèle d'appréciation quantitative des risques est la procédure la plus complexe, mais également la plus importante. Cet article propose quelques développements dans ce domaine ; il est loin d'être exhaustif. Le lecteur intéressé pourra se référer aux ouvrages de Vose [2000] et de Murray [2002], ouvrages accessibles, pratiques et complets dans le domaine.

BIBLIOGRAPHIE

Murray N. ~ Import risk analysis. Animals and animals products. 183 pages, MAF New Zealand, Wellington (NZ), 2002.

Pouillot R., Sanaa M. ~ Bases probabilistes et statistiques nécessaires à l'appréciation du risque. *Epidémiol. et santé anim.*, 2002, **41**, 65-82.

Pouillot R., Sanaa M., Dufour B. ~ Principes de l'appréciation probabiliste quantitative des risques *Epidémiol. et santé anim.*, 2002, **41**, 93-110.

Vose D. ~ Risk analysis. A quantitative guide. 418 pages, Wiley. Chichester, 2000.

REMERCIEMENTS

Je tiens à remercier tout particulièrement Eric Etter* pour l'attention qu'il a portée à la relecture de cet article.

* Agence française de sécurité sanitaire des aliments, Unité d'appui épidémiologique à l'analyse de risque, 27-31 avenue du Gal Leclerc, BP 19, F-94701 Maisons-Alfort Cedex, e.etter@afssa.fr