BASES PROBABILISTES ET STATISTIQUES NECESSAIRES A L'APPRECIATION DU RISQUE

R. Pouillot[1] et M. Sanaa[2]

RESUME : *La réalisation et la compréhension des résultats d'une appréciation des risques nécessitent une connaissance minimale de base dans les domaines de la probabilité et de la statistique. Cet article présente ces quelques notions. Des exemples et des exercices dans le domaine de l'importation d'animaux vivants sont proposés.*

SUMMARY *: The production and the understanding of the results of a risk assessment require some minimal basic knowledge in the field of probability and statistics. This paper presents these observations. Examples and exercises in the field of importing alive animals are proposed.*

Cet article présente quelques bases probabilistes et statistiques nécessaires à la réalisation d'une appréciation du risque. Il ne s'agit que d'un minimum incontournable : il n'est bien entendu pas question d'être exhaustif.

I - NOTIONS DE CALCUL DE PROBABILITES NECESSAIRES EN APPRECIATION QUANTITATIVE DES RISQUES

1. DEFINITIONS

Avant de définir la notion de probabilité, il convient d'évoquer « *l'expérience* » et « *l'événement aléatoire* ».

L'exemple historique le plus classique d'expérience (mais le plus rébarbatif et éloigné de la réalité) est le jet d'un dé non pipé où l'événement attendu, le « numéro de la face supérieure », est qualifié d'aléatoire car dû au hasard. En effet, l'apparition d'une face particulière (pouvant être {1, 2, 3, 4, 5 ou 6}) reste un phénomène non contrôlable.

Des exemples plus pratiques peuvent être proposés dans notre domaine d'intérêt, comme l'expérience consistant à tirer un animal au sort dans une population et à l'associer à l'événement aléatoire « statut infectieux de l'animal » (pouvant être, par exemple, « infecté » ou « indemne »), ou l'expérience consistant à tirer un cheptel au sort dans une population et à l'associer à l'événement aléatoire « taille du cheptel », etc.

[1] Afssa, 27-31 avenue du général Leclerc, BP 19, F-94701 Maisons-Alfort Cedex, France, r.pouillot@afssa.fr.
[2] ENVA, 7 avenue du général de Gaulle, F-94704 Maisons-Alfort cedex, France

Pour ces différents événements, il n'est pas possible de connaître avec certitude le résultat que l'on va obtenir, mais ceci ne signifie pas que l'on ne puisse en avoir aucune connaissance : par exemple, si la maladie est rare dans la population, nous pourrons intuitivement attribuer plus de chances à l'événement « tirage d'un animal indemne » qu'à l'événement « tirage d'un animal infecté ».

A chaque événement possible, résultat d'une expérience, on associe un nombre compris entre 0 et 1, sa probabilité : cette valeur représente une *quantification de la prévision ou la vraisemblance de survenue de l'événement*. De multiples discussions de nature philosophique peuvent être engagées sur la nature « physique », si l'on peut dire, du concept de probabilité [Saporta, 1990]. On retiendra pragmatiquement une définition simple, dite classique : l*a probabilité d'occurrence d'un événement particulier (que l'on notera A) peut être calculée selon la formule* :

$$Pr(A) = \frac{Nombre\ d'événements A}{Nombre\ total\ d'événements possibles}$$

Exemple 1 : Un pays détient 1 000 troupeaux. Le nombre de troupeaux infectés dans ce pays est de 10. Si on tire au sort un troupeau nous aurons 1 000 résultats possibles (un troupeau particulier parmi les 1 000) pour 10 résultats « favorables ». La probabilité pour qu'un cheptel pris au hasard soit infecté est égale à :

Pr(infecté) = 10/1000 = 0,01 = 1%

Exemple 2 : On dispose d'un test de référence qui permet de classer de manière sûre les animaux en « animaux infectés » et « animaux indemnes ». On applique à 100 animaux indemnes et 100 animaux infectés un nouveau test. Le tableau I présente le tableau de contingence obtenu. On a obtenu par exemple 80 animaux infectés et à réponse positive au nouveau test.

Pour obtenir une estimation des divers événements rencontrés dans la population étudiée, il suffit de diviser les valeurs obtenues dans les différentes cases du tableau de contingence par le nombre total d'événements possibles, c'est-à-dire l'effectif total du tableau de contingence. On obtient ainsi les valeurs présentées dans le tableau II. Attention : il s'agit bien d'une évaluation de la probabilité dans la population étudiée. Ainsi, la probabilité estimée d'être infecté de 0,5 ne reflète pas la probabilité d'être infecté dans la population générale puisque, par construction, on avait choisi autant d'animaux infectés que d'animaux indemnes.

TABLEAU I

Tableau de contingence

		Statut		
		Infecté	**Indemne**	**Total**
Nouveau test	Positif	80	10	90
	Négatif	20	90	110
	Total	100	100	200

TABLEAU II

Tableau croisé de probabilités correspondant

		Statut		
		Infecté	**Indemne**	**Total**
Nouveau test	**Positif**	Pr(*Positif et Infecté*) *a* = 80/200 = 0,40	Pr(*Positif et Indemne*) *b* = 10/200 = 0,05	Pr(*Positif*) = *e*= 90/200 = 0,45
	Négatif	Pr(*Négatif et Infecté*) *c* = 20/200 = 0,10	Pr(*Négatif et Indemne*) *d* = 90/200 = 0,45	Pr(Négatif) = *f*= 110/200= 0,55
	Total	Pr(*Infecté*) = *g* = 100/200 = 0,50	Pr(*Indemne*) = *h* = 100/200 = 0,50	1

2. QUELQUES PROPRIETES DES PROBABILITES UTILES EN APPRECIATION QUANTITATIVE DES RISQUES

La probabilité d'un événement certain est 1. La probabilité d'un événement impossible est 0.

PROBABILITE D'UN EVENEMENT CONTRAIRE

Deux événements sont dits « contraires » si les deux événements ne peuvent survenir en même temps et si aucun autre événement ne peut survenir. L'exemple type est l'événement « pile » et l'événement « face », suite au jet d'une pièce de monnaie.

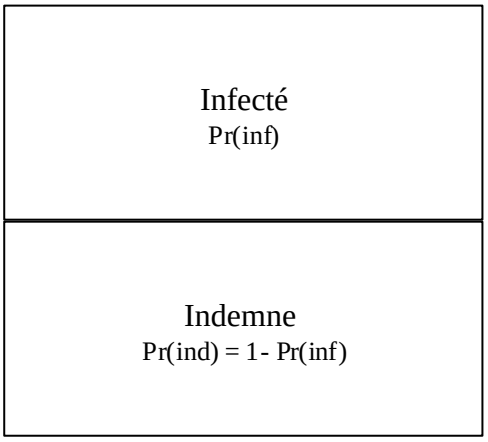
La probabilité de l'événement contraire d'un événement *A*, noté $\overline{A}$ est le complémentaire à 1 de la probabilité de l'événement *A* :

$$\Pr(\overline{A}) = 1 - \Pr(A)$$

Exemple : Pour une maladie donnée, un animal est soit indemne, soit infecté. Il ne peut être les deux à la fois, il ne peut être autre chose. L'événement « être indemne » est le contraire de l'événement « être infecté ». La probabilité qu'un animal soit indemne est le complémentaire à 1 de la probabilité qu'il soit infecté. Si la probabilité d'être infecté est de 0,01, la probabilité d'être indemne est égale à 1 – 0,01 = 0,99 (figure 1).

FIGURE 1

Illustration de deux événements contraires : deux événements sont contraires si, lorsque l'un est réalisé alors l'autre ne l'est pas. En d'autres termes, les deux événements représentent l'univers des possibles, et les deux événements ne peuvent être réalisés en même temps.



PROBABILITE DE L'EVENEMENT (*A* ET *B*) ; PROBABILITE DE L'EVENEMENT (*A* OU *B*)

Dans un premier temps, nous proposerons pour calculer la probabilité de la survenue simultanée de deux événements (événement (*A* et *B*)) l'utilisation de la définition de la probabilité, à savoir :

$$\Pr(A \text{ et } B) = \frac{\text{Nombre d'événements } A \text{ et } B}{\text{Nombre total d'événements possibles}}$$

Nous verrons ultérieurement le calcul de cette probabilité à partir des valeurs Pr(*A*) et Pr(*B*).

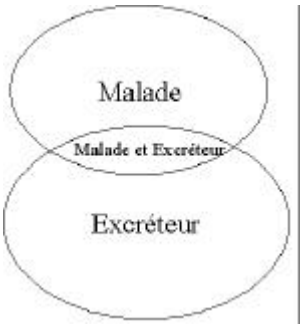
La probabilité de survenue de l'événement (*A* ou *B*) est égale à la somme de la probabilité de A (Pr(*A*)) et de la probabilité de *B* (Pr(*B*)), moins la probabilité de survenue simultanée des deux événements *A* et *B* (Pr(*A* et *B*)) :

$$\Pr(A \text{ ou } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ et } B)$$

Exemple : La probabilité qu'un animal soit malade (c'est-à-dire présente des signes cliniques de la maladie en question) ou excréteur est égale à la somme de la probabilité qu'il soit malade et de la probabilité qu'il soit excréteur, moins la probabilité qu'il soit malade et excréteur (figure 2).

FIGURE 2

La probabilité d'être malade ou excréteur est figurée en pointillés. L'addition simple de la probabilité d'être malade et de la probabilité d'être excréteur reviendrait à compter deux fois la probabilité d'être malade et excréteur.



Si les deux événements ne peuvent avoir lieu en même temps (exemple : « infecté » et « indemne », « testé » et « non testé », « mâle » et « en lactation »,…), les deux événements sont dits exclusifs ou incompatibles. On a alors Pr(*A* et *B*) = 0. On notera donc que la probabilité de l'événement (*A* ou *B*) est égale à la somme des probabilités de A et de B si A et B sont incompatibles :

Pr(*A* ou *B*) = Pr(*A*) + Pr(*B*)
si *A* et *B* sont incompatibles

Conclusion : on ne peut additionner des probabilités pour évaluer la probabilité de survenue d'un événement ou d'un autre que si les événements sont incompatibles.

Exemple : Si la probabilité que l'animal soit cliniquement malade est de Pr(malade) = 0,6, la probabilité que l'animal soit excréteur est Pr(excréteur) = 0,6 et la probabilité que l'animal soit malade et excréteur est Pr(malade et excréteur) = 0,3, la probabilité que l'animal soit malade ou excréteur est :

Pr(*malade* ou *excréteur*) =
Pr(*malade*) + Pr(*excréteur*)
– Pr(*malade* et *excréteur*) =
0,6 + 0,6 – 0,3 = 0,9

On notera que la simple addition des probabilités Pr(malade) et Pr(excréteur) aboutirait à un résultat de 1,2, ce qui est impossible.

Exemple : La probabilité qu'un animal fournisse une réponse négative à un test donné est de 95%. La probabilité qu'un animal ne soit pas testé est de 20%. En pratique, l'animal est introduit sur le territoire si la réponse au test est négative ou si, pour diverses raisons, il n'est pas testé. La probabilité que l'animal soit introduit est égale à la somme de la probabilité que la réponse au test soit négative et de la probabilité qu'il ne soit pas testé, moins la probabilité qu'il soit non testé et (potentiellement) négatif :

Pr(*Introduit*) = Pr(*Négatif*) + Pr(*Non testé*)
– Pr(*Négatif* et *Non testé*)

PROBABILITES DE L'EVENEMENT (*A* si *B*)

Dans différentes situations pratiques, il est nécessaire de connaître la probabilité d'un événement conditionnellement à un autre événement, c'est-à-dire la probabilité d'un événement sachant qu'un autre événement est survenu. C'est le cas par exemple, de la probabilité d'excrétion en fonction du statut de l'animal. Il peut être nécessaire, dans certains cas, de connaître la probabilité de l'excrétion sachant que l'animal est infecté de façon asymptomatique, et non sur l'ensemble des animaux infectés. La probabilité de *A* sachant *B*, que l'on appelle également « probabilité de *A* si *B* » ou « probabilité de *A* conditionnellement à *B* » et que l'on écrit Pr(*A*|*B*) est égale à :

Equation (1) $$\Pr(A|B) = \frac{\Pr(A \text{ et } B)}{\Pr(B)}$$

Cette équation s'écrit de manière équivalente :

Equation (2) Pr(*A* et *B*) = Pr(*A* | *B*) Pr(*B*)

Cette équation revient à calculer :

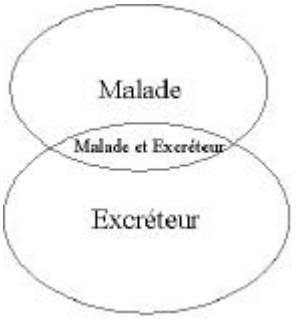
$$\Pr(A|B) = \frac{\text{Nombre d'événements}(A \text{ et } B)}{\text{Nombre d'événements } B}$$

Exemple : On s'intéresse à la probabilité d'être excréteur sachant que l'on est malade. Dans la

figure 3, on ne s'intéresse donc plus à tous les animaux (malades et excréteurs), mais uniquement aux animaux malades. On désire donc évaluer la probabilité d'être malade parmi les excréteurs. On confirme graphiquement que cette valeur est bien égale à la probabilité d'être malade et excréteur divisée par la probabilité d'être malade.

FIGURE 3

La probabilité d'être excréteur sachant que l'on est malade est le rapport de la surface en pointillés (malade et excréteur) sur la surface ovale supérieure (malade).



Exemple : La sensibilité d'un test sérologique est la probabilité que le test soit positif (événement *A*) si l'animal est infecté (événement *B*). Cette sensibilité est égale à la probabilité que l'animal soit infecté et positif au test divisé par la probabilité que l'animal soit infecté. On a estimé le tableau de probabilités (cf. tableau 2) des différents événements possibles. La sensibilité du test est évaluée selon :

$$\Pr(\text{Positif}|\text{Infecté}) = \frac{\Pr(\text{Positif et Infecté})}{\Pr(\text{Infecté})} = \frac{a}{g} = \frac{0,4}{0,5} = 0,8$$

INDEPENDANCE, PROBABILITE DE L'EVENEMENT (*A* et *B*)

On dit que *A* est indépendant de *B* si la probabilité de l'événement *A* est la même quelle que soit la réalisation de l'événement *B*. Autrement dit, la connaissance d'un événement ne change pas les chances de réalisation de l'autre :

$$\Pr(A|B) = \Pr(A) \text{ quel que soit } B$$

La compréhension de cette notion est essentielle en appréciation des risques.

La dépendance peut être très intuitive. Par exemple, la probabilité qu'un animal soit positif à un test sachant qu'il est infecté ne sera pas la même que la probabilité que l'animal soit positif au test sachant qu'il est indemne : résultat du test et statut infectieux ne sont (heureusement) pas indépendants. Dans d'autres cas, la dépendance est plus indirecte : un exemple pourrait être le résultat d'un test histologique réalisé par deux anatomo-pathologistes « en aveugle » (c'est-à-dire un examinateur ne connaît pas le résultat obtenu par l'autre examinateur). Malgré l'intuition d'une indépendance des résultats (liée à l'indépendance de la réalisation de l'analyse), la probabilité de trouver une lame positive par le second observateur sera généralement plus importante si le premier observateur a trouvé cette lame positive que si le premier observateur l'a trouvée négative.

Conséquence de l'indépendance : On note, en réarrangeant l'équation (2) et si les deux événements sont indépendants :

Equation (3) A et B indépendants $\Rightarrow$
$\Pr(A \text{ et } B) = \Pr(A)\ \Pr(B)$

Si deux événements sont indépendants, la probabilité d'obtenir en même temps (conjointement) les deux événements est égale au produit des probabilités de chaque événement. Dans la construction de l'arbre de probabilité [Toma, 2002], on sera souvent amené à évaluer la probabilité de survenue conjointe ou la survenue conditionnelle de deux événements. Un raccourci rapide intuitif consisterait à multiplier les probabilités. Or, ceci n'est possible que si les événements sont indépendants.

Conclusion : on ne peut multiplier des probabilités pour évaluer la probabilité conjointe de deux événements qu'après s'être assuré de l'indépendance de ces événements.

Les événements suivants sont souvent considérés comme indépendants, mais cette indépendance doit être vérifiée et discutée au cas par cas :

- résultats de deux tests de diagnostic : très souvent, les résultats sont considérés comme indépendants. Mais en pratique, notamment si les tests sont fondés sur la même base physiologique (par exemple exploration d'un même type de réaction immunologique), un animal infecté et positif à un test a une probabilité plus grande d'être positif à un second test qu'un animal infecté sans connaissance du résultat du premier test. Il n'y a souvent pas d'indépendance entre les résultats de tests diagnostiques ;
- probabilité d'infection et sexe (ou âge) des animaux : la probabilité d'infection est souvent variable selon le sexe ou l'âge des animaux, ceci pour des raisons physiologiques (exemple évident : mammites), pour des raisons liées à la dynamique de la maladie (les jeunes animaux peuvent être plus touchés que les adultes car nouvellement exposés au risque), pour des raisons de modes de transmission ou encore des raisons de techniques d'élevage (les jeunes sont plus confinés).

Pour ces types d'événements, une prise en compte de la dépendance, ou au moins une discussion de l'hypothèse d'indépendance, devra être réalisée. L'indépendance entre deux événements peut être explorée empiriquement grâce à l'observation simultanée des deux critères. Si, dans un tableau croisé, on observe que $\Pr(A \text{ et } B) \neq \Pr(A) \times \Pr(B)$, il faudra considérer les événements comme non indépendants.

Exemple : Dans le tableau II, on a :

$\Pr(\textit{positif} \text{ et } \textit{infecté}) = a = 0{,}40$ et $\Pr(\textit{positif}) \times \Pr(\textit{infecté}) = e \times g = 0{,}45 \times 0{,}50 = 0{,}225$

Le résultat du test sérologique et le statut infectieux ne sont, bien évidemment, pas indépendants[3].

Exemple : Reprenons l'exemple précédent : la probabilité qu'un animal fournisse une réponse négative à un test donné est de 95%. La probabilité qu'un animal ne soit pas testé est de 20%. La probabilité d'introduire un animal infecté s'exprime selon :

$$\Pr(\textit{Introduit}) = \Pr(\textit{Négatif}) + \Pr(\textit{Non testé}) - \Pr(\textit{Négatif} \text{ et } \textit{Non testé})$$

Si l'on considère que le fait de ne pas être testé est indépendant du résultat qu'aurait obtenu l'animal, on a :

$$\Pr(\textit{Négatif} \text{ et } \textit{Non testé}) = \Pr(\textit{Négatif}) \times \Pr(\textit{Non testé})$$

On a donc :

$$\Pr(\textit{Introduit}) = 0{,}95 + 0{,}20 - 0{,}95 \times 0{,}20 = 96\%$$

La question se pose de savoir si le fait de ne pas être testé est indépendant du résultat qu'aurait obtenu l'animal : cette indépendance semble intuitive, mais il est possible d'envisager des cas de figures différents : fraude permettant aux animaux infectés d'éviter le test, évitement du test par une sous-population d'âge (ou de sexe, ou de région géographique,…) moins (ou plus) infectée que la moyenne,…

3. THEOREME DE BAYES

Le théorème de Bayes permet, connaissant la probabilité de *A* sachant *B*, de calculer la probabilité de *B* sachant *A*. Il s'écrit :

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + \Pr(A|\overline{B})\Pr(\overline{B})}$$

Exemple : Un animal provient d'une zone dans laquelle la prévalence animale d'une maladie est égale à 1%. L'animal subit un test sérologique de sensibilité Se = 0,8 et de spécificité Sp = 0,99. Il est positif à ce test. Quelle est la probabilité que l'animal soit infecté ?

On peut exprimer les données ainsi : on cherche la probabilité d'être malade sachant que le test est positif soit $\Pr(\textit{malade}|\textit{positif})$.

3 Nota : pour estimer le caractère indépendant de deux événements, il est possible de faire un test classique du χ^2 sur le tableau croisé de fréquences observées de deux événements. Si le test est significatif, les

événements ne sont probablement pas indépendants; si le test n'est pas significatif (et que la dépendance n'est pas intuitive), on peut considérer que les événements sont indépendants.

On dispose de la sensibilité, qui n'est autre que la probabilité d'être positif si l'animal est malade soit Pr(*positif*|*malade*), la spécificité, qui n'est autre que la probabilité d'être négatif si l'animal est indemne soit Pr(*négatif*|*indemne*), soit :

$$1 - \Pr(positif|indemne)$$

(car les événements « être positif au test » et « être négatif au test » sont contraires), et la prévalence n'est autre que Pr(*malade*). On a, selon le théorème de Bayes :

$$\Pr(malade|positif) = \frac{\Pr(positif|malade)\Pr(malade)}{\Pr(positif|malade)\Pr(malade) + \Pr(positif|indemne)\Pr(indemne)}$$

$$= \frac{Se\ p}{Se\ p + (1 - Sp)(1 - p)}$$

On aura reconnu la formule de la valeur prédictive positive.

Le théorème de Bayes permet d'actualiser une probabilité a priori (la probabilité d'infection, c'est-à-dire la prévalence), en fonction d'une donnée observée (un test positif). Il est à la base de toute une approche statistique dite « Bayésienne ».

II - NOTIONS DE STATISTIQUE NECESSAIRES EN APPRECIATION QUANTITATIVE DES RISQUES

Soit une variable résultat d'un tirage au sort (par exemple le nombre d'animaux infectés issus d'un tirage au sort dans une population, le poids d'un animal pris au sort dans une population,….). On appellera ce type de variable « variable aléatoire », et on la notera *X*.

1. LOI DE PROBABILITE

CAS DISCRET

Si *X* ne peut prendre qu'un certain nombre de valeurs, cette variable est dite discrète (exemple : nombre d'animaux infectés par troupeau). La loi de probabilité de *X* est la fonction associant à toute valeur *x* possible de *X* la probabilité associée Pr(*X* = x). La fonction de répartition de la variable aléatoire est la fonction associant à toute valeur x la probabilité que la valeur soit inférieure ou égale à x : F(x) = Pr(X ≤ x).

Exemple : On soumet cinq animaux infectés à un test sérologique de sensibilité 40%. Le nombre d'animaux à réponse positive au test peut varier de 0 à 5. Le tableau III présente la loi de probabilité et la fonction de répartition du nombre d'animaux à réponse positive au test. La probabilité d'obtenir cinq animaux à réponse positive est de 1%, la probabilité d'obtenir trois animaux à réponse positive ou moins est 91%,…

TABLEAU III

Exemple de loi de distribution discrète

Nombre d'animaux à réponse positive	Loi de probabilité Pr(*X* = x)	Fonction de répartition Pr(*X* ≤ x)
0	8%	8%
1	26%	34%
2	35%	68%
3	23%	91%
4	8%	99%

5	1%	100%

CAS CONTINU

Si X peut prendre un nombre infini de valeurs, cette variable est dite continue. La fonction de répartition est encore la fonction associant à toute valeur x la probabilité $\Pr(X \leq x)$. La loi de probabilité de X , notée f(x) n'a pas de correspondance probabiliste immédiate : il s'agit de la dérivée de la fonction de répartition par rapport à x. Ainsi, f(x) peut être supérieure à 1.

Des exemples de loi de distribution et de fonction de répartition de variables discrètes et continues sont proposés figure 4.

3. ESPERANCE, MOYENNE

Une loi de probabilité peut être caractérisée par des valeurs particulières, que nous appellerons « statistiques ».

L'espérance représente le « centre de gravité » de la distribution. Il s'agit d'une valeur exacte, estimée par la moyenne[4]. On notera l'espérance d'une variable aléatoire X : E(X)

Définition

Dans le cas discret, on a :

$$E(X) = \sum_{x} x \Pr(X = x),$$

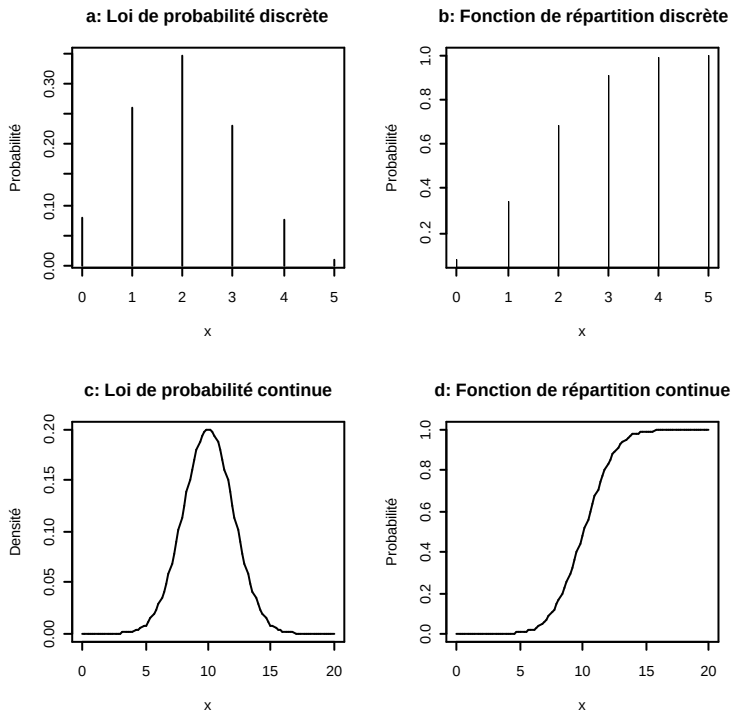
x représentant l'ensemble des valeurs possibles de x

Dans le cas continu, on a :

$$E(X) = \int_{x=-\infty}^{+\infty} x\, f(x)$$

FIGURE 4

Loi de probabilité d'une variable discrète (a) : à chaque valeur possible de x (en abscisse), on peut associer sa probabilité (en ordonnée) ; Fonction de répartition d'une variable discrète (b) : ces deux fonctions sont représentées sous forme d'histogramme ; Loi de probabilité d'une variable continue (c) ; Fonction de répartition d'une variable continue (d) : ces deux fonctions sont représentées sous forme de courbe.



[4] En pratique, ces deux termes sont bien souvent confondus. On les emploiera indistinctement dans le reste de cet article.

Propriétés

Ces propriétés sont fondamentales en appréciation quantitative des risques : elles sont notamment le fondement de l'estimation « ponctuelle » de risque. On a, si X et Y sont deux variables aléatoires, a et b sont deux constantes :

$$E(a\,X + b) = a\,E(X) + b$$

$E(X + Y) = E(X) + E(Y)$, quelles que soient X et Y.

La moyenne de la somme de variables aléatoires est égale à la somme des moyennes.

Exemple (trivial) : Si l'on importe en moyenne 30 animaux de deux pays différents, on importe en moyenne 60 animaux en tout.

$E(XY) = E(X)\,E(Y)$ seulement si X et Y sont indépendantes.

La moyenne du produit de variables aléatoires indépendantes est égale au produit des moyennes.

Exemple (trivial) : Si l'on importe en moyenne 100 animaux et que, en moyenne, 1% des animaux est infecté, on importe en moyenne un animal infecté.

Contre-exemple : Si le coût d'une maladie est égal au carré du nombre d'animaux infectés, et que l'on a en moyenne 100 animaux infectés, le coût moyen n'est pas égal à 10 000 (car $E(X^2) = E(X \times X) \neq E(X) \times E(X)$ car X et X ne sont, par définition, pas indépendants.

Notons également que :

$$E\left(\frac{1}{Y}\right) \neq \frac{1}{E(Y)}$$

et, par conséquent,

$$E\left(\frac{X}{Y}\right) \neq \frac{E(X)}{E(Y)}$$

La moyenne du rapport de variables aléatoires n'est pas égale au rapport des moyennes.

Estimation de l'espérance

L'estimation de l'espérance est la moyenne, notée m(X). Si l'on dispose d'une série de données, on a :

$$m(x_1,\ldots,x_n) = \frac{\sum_{i=1}^{n} x_i}{n}$$

C'est-à-dire la somme des données divisées par le nombre de données.

Ecriture Excel© de la moyenne d'une série de données

Si les données sont situées dans les cellules A2 à A20, le calcul de la moyenne est réalisé en appliquant la formule : « =MOYENNE(A2:A20) » (cf. figure 5).

3. VARIANCE, ECART-TYPE

La variance est une statistique de dispersion : plus les données sont dispersées, plus sa valeur est élevée. On notera la variance d'une variable aléatoire X : V(X). L'unité d'expression de la variance est l'unité de la variable élevée au carré. L'écart-type est la racine carrée de la variance : il a pour avantage d'être exprimé dans la même unité que la variable aléatoire. On notera l'écart-type d'une variable aléatoire X : σ(X).

Définition :

On a :

$$V(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

Propriétés

On a, si X et Y sont deux variables aléatoires, a et b sont deux constantes :

$$V(a\,X) = a^2\,V(X)$$

$$V(X + b) = V(X)$$

On a donc, en ce qui concerne les écart-types :

Equation (4) $s(a\,X) = a\;s(X)$

et $s(X + b) = s(X)$.

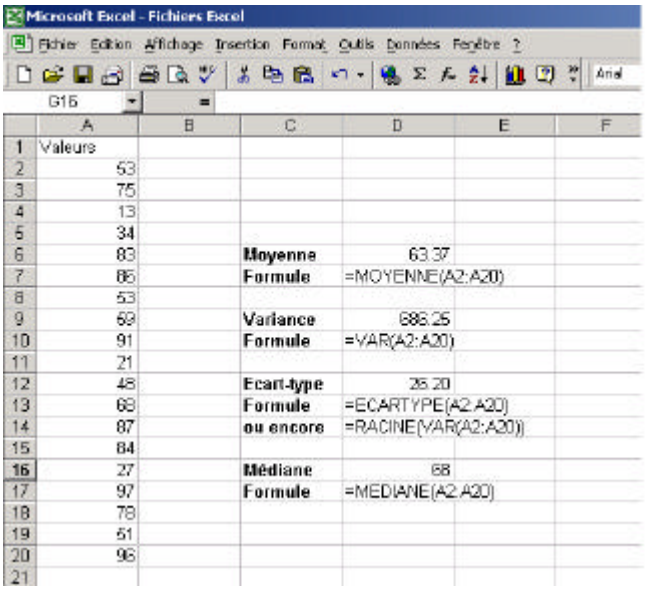
Exemple : Tous les ans, on importe un nombre variable d'animaux d'un pays donné. Chaque animal importé subit un test diagnostic coûtant cinq euros. La variance du nombre d'animaux importés est de 100. La variance du coût induit par ces achats est donc $V(5\,n) = 25\,V(n) = 2\,500$ euros.

Equation (5) $V(X + Y) = V(X) + V(Y)$, seulement si X et Y sont indépendantes[5].

[5] Si X et Y ne sont pas indépendants, on a $V(X + Y) = V(X) + V(Y) + 2\ cov(X,Y)$ où $cov(X,Y) = E(XY) – E(X)\ E(Y) = E[(X – E(X))(Y – E(Y))]$ est la covariance de X et Y.

FIGURE 5

Présentation d'une feuille de calcul Excel© des principales statistiques



La variance de la somme de variables aléatoires indépendantes est égale à la somme des variances.

D'où l'on déduit :

$V(X - Y) = V(X) + V(Y)$, seulement si X et Y sont indépendantes.

La variance de la différence de variables aléatoires indépendantes est égale à la somme des variances.

On a donc, en ce qui concerne les écart-types :

$\sigma(X + Y) = \sqrt{\sigma(X)^2 + \sigma(Y)^2}$ seulement si X et Y sont indépendantes.

$\sigma(X - Y) = \sqrt{\sigma(X)^2 + \sigma(Y)^2}$ seulement si X et Y sont indépendantes.

Exemple 1 : On importe indépendamment, d'année en année, des animaux de deux pays différents. Si l'écart-type du nombre d'animaux importés par pays est 5, l'écart-type du nombre total d'animaux importés est $\sqrt{5^2 + 5^2} = 7{,}07$.

Exemple 2: On importe, d'année en année, des animaux de deux pays différents. Chaque année, on importe autant d'animaux d'un pays que de l'autre. Si l'écart-type du nombre d'animaux importés par an d'un pays est 5, l'écart-type du nombre total d'animaux importés est 2 x 5 = 10. En effet, on importe chaque année $X + X$ animaux : il faut donc utiliser l'équation (4) : $\sigma(2X) = 2\,\sigma(X)$ et non l'équation (5) : il y a totale dépendance entre les variables.

Estimation de la variance

Si l'on dispose d'une série de données, on a :

$$\mathrm{var}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n-1} = \frac{n\sum_{i=1}^{n} x_i^{\,2} - m^2}{n(n-1)}$$

Ecriture Excel© de l'estimation de la variance d'une série de données

Si les données sont situées dans les cellules A2 à A8, le calcul de la moyenne est réalisé en appliquant la formule : « =VAR(A2:A8) » (cf. figure 5).

4. MEDIANE ET PERCENTILES

La médiane est la valeur de la variable aléatoire telle que la moitié (50%) des observations seront inférieures, la moitié (50%) des observations seront supérieures. Notons que la médiane ne doit pas être confondue avec la moyenne (cf. figure 6).

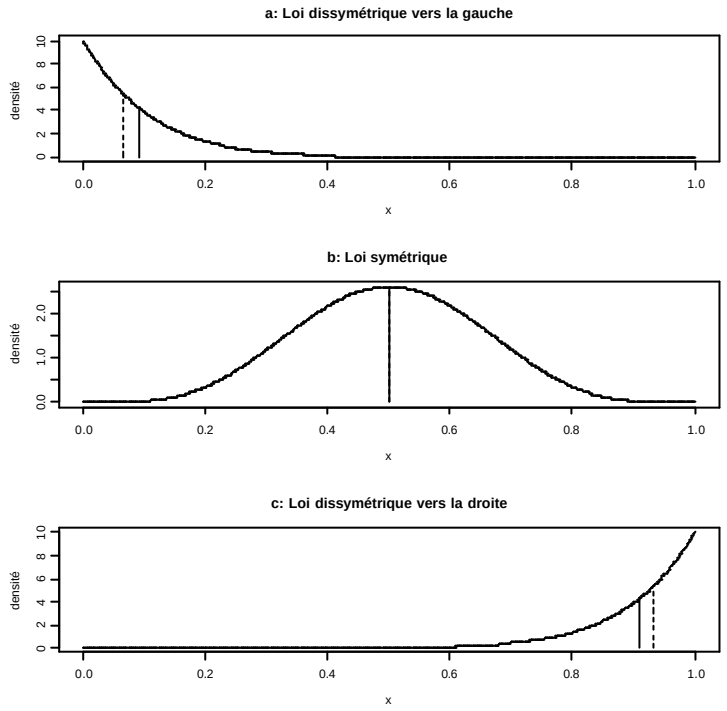
Ecriture Excel© de l'estimation de la médiane d'une série de données

Si les données sont situées dans les cellules A2 à A20, le calcul de la moyenne est réalisé en appliquant la formule : « =MEDIANE(A2:A20) » (cf. figure 5).

Le cinquième percentile est la valeur telle que 5% des observations seront inférieures, 95% seront supérieures. Plus généralement, le $p^{ème}$ percentile sera la valeur telle que p% des observations seront inférieures, (100 - p) % seront supérieures.

FIGURE 6

Si la loi de distribution est dissymétrique vers la gauche (a), la moyenne (trait plein) est supérieure à la médiane (trait pointillé) ; si la loi de distribution est symétrique (b), moyenne et médiane sont confondues; si la loi de distribution est dissymétrique vers la droite (c), la moyenne est inférieure à la médiane.



III - CONCLUSION

L'appréciation quantitative des risques repose sur des calculs probabilistes et statistiques. La compréhension, la réalisation et la discussion de ces analyses ne peuvent être valables que si l'on connaît les bases probabilistes et statistiques utilisées. Par exemple, si « l'évaluateur » multiplie des probabilités lors de son analyse, il faut savoir repérer l'hypothèse sous-jacente d'une indépendance entre les événements dont on multiplie les probabilités et vérifier que cette hypothèse est valable et/ou discutée.

Ces quelques notions élémentaires, notamment les propriétés fondamentales des probabilités et des deux principales statistiques (espérance et variance), devraient permettre de comprendre la plupart des problèmes d'appréciation quantitative du risque, au moins pour les évaluations « ponctuelles ».

Les notions développées dans ce texte sont présentes dans tout ouvrage élémentaire traitant de probabilité et de statistique. Le lecteur intéressé est invité à se référer à ce type d'ouvrage pour obtenir des compléments. On retiendra notamment :

Lazar Ph. - Eléments de probabilité et statistique, 1998.

ou un ouvrage plus complet (donc complexe) tel que :

Saporta G. - Probabilités, analyse de données et statistique. 493 pages. Editions TECHNIP. Paris. 1990.

On retiendra également l'ouvrage :

Schwartz D. - Le jeu de la science et du hasard, Flammarion, Paris, 1994. pour une présentation très imagée et accessible de la théorie et du mode de pensée statistique.

BIBLIOGRAPHIE

Saporta G. ~ Probabilités, analyse de données et statistique. 493 pages. Editions TECHNIP. Paris. 1990.

Toma B. ~ L'appréciation quantitative du risque : notions générales. *Epidémiol. et santé anim*., 2002, ce numéro.

Toma B., Dufour B., Sanaa M., Bénet J.J., Shaw A., Moutou F., Louza A. ~ Epidémiologie appliquée à la lutte collective contre les maladies animales transmissibles majeures (2e édition). 696 pages, Association pour l'étude de l'épidémiologie des maladies animales, Maisons Alfort, 2001.

ANNEXE

EXERCICE 1

Connaissant la prévalence p d'une maladie dans une population bovine, la sensibilité Se d'un test, et la spécificité Sp de ce test, calculer :

a) la probabilité qu'une vache soit indemne ;

b) la probabilité qu'une vache infectée soit négative au test ;

c) la probabilité qu'une vache indemne soit positive au test ;

d) la probabilité qu'une vache soit infectée et positive au test ;

e) la probabilité qu'une vache soit indemne et négative au test ;

f) la probabilité qu'une vache soit infectée et négative au test ;

g) la probabilité qu'une vache soit indemne et positive au test ;

h) la probabilité qu'un vache soit positive au test ;

i) la probabilité qu'un vache soit négative au test ;

j) la probabilité qu'une vache positive au test soit infectée ;

k) la probabilité qu'une vache négative au test soit indemne ;

l) la probabilité qu'une vache positive au test soit indemne ;

m) la probabilité qu'une vache négative au test soit infectée.

EXERCICE 2

On désire importer n animaux d'une zone où la prévalence d'une maladie est p.

a) quelle est la probabilité d'importer au moins un animal infecté ?

On teste les animaux à l'aide d'un test de sensibilité Se et de spécificité Sp.

b) quelle est la probabilité d'observer au moins un animal positif ?

Les animaux positifs ne sont pas importés ; les animaux négatifs sont importés.

c) quelle est la probabilité qu'au moins un animal infecté soit importé ?

Application numérique : p = 0,1%, Se = 80%, Sp= 99% , n = 100.

SOLUTION DE L'EXERCICE 1 :

Il est nécessaire de traduire les données en terme de probabilité.

- la prévalence est p = Pr(*inf*) : c'est la probabilité d'être infecté ;
- la sensibilité est Se = Pr(+|*inf*) : c'est la probabilité d'être positif si l'animal est infecté ;
- la spécificité est Sp = Pr(-|*ind*)) : c'est la probabilité d'être négatif si l'animal est indemne.

D'autre part, on reconnaît les événements contraires : « Infecté » et « Indemne » d'une part, « Négatif » et « Positif » d'autre part.

a) la probabilité qu'une vache soit indemne = 1 - p

Une vache est soit indemne, soit infectée : ces deux événements sont contraires. On a donc : Pr(*ind*) = 1 - Pr(*inf*) = 1-*p*

b) la probabilité qu'une vache infectée soit négative au test = 1-*Se*

Dans la sous-population des vaches infectées, les vaches sont soit positives, soit négatives au test : dans la sous-population des vaches infectées, ces événements sont donc contraires. On a donc : Pr(-|inf) = 1 - Pr(+|inf) = 1-Se

c) la probabilité qu'une vache indemne soit positive au test = 1-*Sp*

Avec le même raisonnement dans la population indemne, on a : Pr(+|*ind*) = 1 - Pr(-|*ind*) = 1-*Sp*

d) la probabilité qu'une vache soit infectée et positive au test = *p Se*

Selon l'0, on peut écrire Pr(*inf et* +) = Pr(*inf*) Pr(+|*inf*) =*pSe*.

e) la probabilité qu'une vache soit indemne et négative au test = (1-*p*) *Sp*

Selon l'0, on peut écrire Pr(*ind et* -) = Pr(*ind*) Pr(-|*ind*). Or on a Pr(*ind*) = 1-p (solution du a), d'où Pr(*ind* et -) = (1-p) *Sp*

f) la probabilité qu'une vache soit infectée et négative au test = p (1-*Se*)

Selon l'0, on peut écrire Pr(*inf* et -) = Pr(*inf*) Pr(-|*inf*). Or on a Pr(-|*inf*) = 1-Se (solution du b), d'où Pr(*inf* et -) = p *Se*

g) la probabilité qu'une vache soit indemne et positive au test = (1-p)(1- *Sp*)

Selon l'0, on peut écrire Pr(*ind* et +) = Pr(*ind*) Pr(+|*ind*). Or on a Pr(*ind*) = 1-p (solution du a) et Pr(+|*ind*) = 1-*Sp* (solution du c), , d'où Pr(*ind* et +) = (1-p)(1- *Sp*)

h) la probabilité qu'un vache soit positive au test = p Se + (1-p) (1-Sp)

Il existe deux possibilités pour qu'une vache soit positive au test : soit elle est infectée et positive, soit elle est indemne et positive. On a Pr(+) = Pr(+|inf) + Pr(+|*ind*). Selon les solutions du (d) et du (g), on en déduit la solution Pr(+) = p Se + (1-p) (1- *Sp*)

i) la probabilité qu'un vache soit négative au test = p (1- *Se*) + (1-p) *Sp*

De la même façon, Il existe deux possibilités pour qu'une vache soit négative au test : soit elle est infectée et négative (solution du f), soit elle est indemne et négative (solution du e) g), on en déduit la solution Pr(-) = p (1-Se) + (1-p)Sp

j) la probabilité qu'une vache positive au test soit infectée = $\frac{p\ Se}{p\ Se+(1-p)(1-Sp)}$.

On applique le théorème de Bayes :

$$\Pr(\text{inf}\,|+)=\frac{\Pr(+|\text{inf})\Pr(\text{inf})}{\Pr(+|\text{inf})\Pr(\text{inf})+\Pr(+|ind)\Pr(ind)}=\frac{Se\ p}{Se\ p+(1-Sp)(1-p)}$$

k) la probabilité qu'une vache négative au test soit indemne = $\frac{Sp(1-p)}{Sp(1-p)+(1-Se)p}$

On applique le théorème de Bayes :

$$\Pr(ind\,|-)=\frac{\Pr(-|ind)\Pr(ind)}{\Pr(-|ind)\Pr(ind)+\Pr(-|\text{inf})\Pr(\text{inf})}=\frac{Sp(1-p)}{Sp(1-p)+(1-Se)p}$$

l) la probabilité qu'une vache positive au test soit indemne = $\frac{(1-Sp)(1-p)}{(1-Sp)(1-p)+\ p\ Se}$

On applique le théorème de Bayes :

$$\Pr(ind|+) = \frac{\Pr(+|ind)\Pr(ind)}{\Pr(+|ind)\Pr(ind) + \Pr(+|\inf)\Pr(\inf)} = \frac{(1-Sp)(1-p)}{(1-Sp)(1-p) + Se\ p}$$

m) la probabilité qu'une vache négative au test soit infectée = $\frac{(1-Se)p}{(1-Se)p + Sp(1-p)}$

On applique le théorème de Bayes :

$$\Pr(\inf|-) = \frac{\Pr(-|\inf)\Pr(\inf)}{\Pr(-|\inf)\Pr(\inf) + \Pr(-|ind)\Pr(ind)} = \frac{(1-Se)p}{(1-Se)p + Sp(1-p)}$$

SOLUTION DE L'EXERCICE 2 :

a) Il serait nécessaire d'évaluer la probabilité d'importer un animal infecté, deux animaux infectés,…, n animaux infectés, puis de faire la somme de ces probabilités (car on importe 1 ou 2 ou 3 ou…n animaux infectés). Il est plus facile de raisonner comme suit : soit on importe 0 animal infecté, soit on importe au moins un animal infecté. Ces événements sont contraires. On a donc Pr(au moins 1 inf) = 1-Pr(0 inf).

Si la prévalence dans la population est p, la probabilité pour un animal de ne pas être infecté est (1-p) (événements contraires).

La probabilité que le premier animal importé soit indemne est donc (1-p) ;

La probabilité que les deux premiers animaux importés soient indemnes est

Pr(1° animal indemne) $\times$ Pr(2° animal indemne) = (1-p) $\times$ (1-p) = $(1-p)^2$

(selon l'0, si les événements sont indépendants).

La probabilité que les trois premiers animaux importés soient indemnes est :

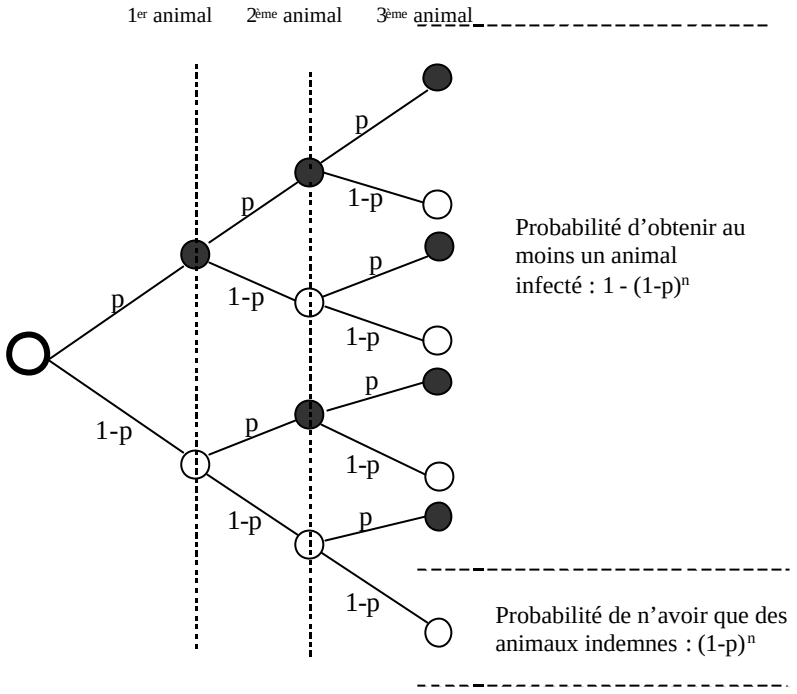
Pr(1° animal indemne) $\times$ Pr(2° animal indemne) $\times$ Pr(3° animal indemne) =
(1-p) $\times$ (1-p) $\times$ (1-p) = $(1-p)^3$

…

Ce raisonnement peut, plus intuitivement, être présenté sous la forme d'un arbre de probabilités (cf. figure 7).

FIGURE 7

Schéma représentant les probabilités de résultats lors de tirage au sort d'animaux
[Toma *et al.*, 2001]



La probabilité que les *n* animaux importés soient indemnes est donc $(1-p)^n$

La solution est donc Pr(au moins 1 *inf*) = 1-Pr(0 *inf*) = $1-(1-p)^n$.

Une question se pose : les tirages au sort sont-ils indépendants ? La probabilité que le deuxième animal soit indemne est-elle influencée par le statut infectieux du premier animal tiré au sort ? Cette question peut également se poser de la manière suivante : la prévalence varie-t-elle lorsque l'on retire des animaux de la population ?. La réponse est : oui. Rappelons que la prévalence est le rapport (Nb d'animaux infectés dans la population)/(Nb total d'animaux dans la population). Alors, si notre premier animal est infecté, le nombre d'animaux infectés restant dans la population est plus faible de une unité, et le nombre total d'animaux restant dans la population diminue également de une unité : la prévalence diminue. Si le premier animal est indemne, le nombre d'animaux infectés restant dans la population est inchangé, et le nombre total d'animaux restant dans la population diminue de une unité : la prévalence augmente… En pratique, cependant, ces variations sont négligeables si l'échantillon est faible au regard de la population totale : le retrait d'un animal ne modifie alors que très faiblement la prévalence… (notion de population finie ou infinie et le seuil n/N inférieur à 10%) On peut donc considérer les tirages au sort comme indépendants, sous réserve que l'échantillon soit petit au regard de la population générale.

Cette digression a pour seul but de montrer qu'il est nécessaire de réfléchir avant d'employer des formules type « recette ».

Application numérique : $1-(1-0{,}001)^{100} = 10\%$

b) Selon le même raisonnement, deux possibilités contraires existent : soit tous les animaux sont négatifs, soit au moins un animal est positif.

Un animal est négatif selon deux modalités mutuellement exclusives (arbre d'événement) :

- soit il est infecté et négatif ;
- soit il est indemne et négatif.

La probabilité qu'il soit négatif est donc (Solution de l'exercice 1i) : p (1-Se) + Sp(1-p)

Selon le même raisonnement que le a), la probabilité que tous les animaux soient négatifs est donc : (p (1-Se) + Sp(1-p))^n ;

La probabilité qu'au moins un animal soit positif est donc 1-(p (1-Se) + Sp(1-p))^n

Application numérique : 1-(0,001 (1-0,8) + 0,99(1-0,001))^{100} = 10%

c) Il existe deux événements contraires (arbre d'événement) :

- soit l'animal est infecté et non détecté : il y a alors importation de la maladie. La probabilité de cet événement est p(1-Se) (exercice 1f) ;
- soit l'animal est indemne : il ne pose donc pas de problème, soit il est infecté et détecté : il ne pose également pas de problème. La probabilité de cet événement, contraire du précédent, est 1-p(1-Se).

Selon le même raisonnement que précédemment, la probabilité que tous les animaux ne posent pas de problème est donc (1-p(1-Se))^n.

La probabilité pour qu'il y ait un problème est alors 1-(1-p(1-Se))^n.

Application numérique : 1-(1-0,001 (1-0,8))^{100} = 2%