La multiplication spontanée des données.

Claudine Schwartz

L'article de B. Laponche et B. Dessus dans le journal Libération (5 juin 2011) énoncant que la probabilité d'avoir un accident nucléaire grave dans les 30 années à venir était 50% (!!) pour la France et de plus de 100% (????) pour l'Europe a donné lieu à divers commentaires[1]. Des questions m'ont été posées que j'ai regroupées ci-dessous.

1- Les apparences sont trompeuses

Si p est la probabilité de succès d'une épreuve dont le résultat est 0 (échec) ou 1(succès), et si p est petit, la probabilité p' d'avoir au moins un succès sur n épreuves (n pas trop grand) est environ np. Ce produit de la probabilité par le nombre de répétitions est trompeur *:* il ne résulte pas d'un raisonnement qui justifierait de multiplier p par n, mais d'une approximation numérique que l'on peut faire pour certaines valeurs (petites) de p et (pas grandes) de n, puisqu'en fait $p'=1-(1-p)^n$.

Avec p=0,0003, le tableau ci-dessous montre illustre comment se dégrade la qualité de l'approximation de p' par np lorsque n croit[2].

n	$1-(1-0{,}0003)^n$	$0{,}0003n$
1	0,0003	0,0003
4	0,0012	0,0012
16	0,0048	0,0048
64	0,0190	0,0192
256	0,0739	0,0768
1024	0,2645	0,3072
4096	0,7074	1,2288

2-La multiplication spontanée des observations

On observe le fonctionnement de 450 réacteurs nucléaires pendant 30 années et on trouve R pannes. Faisons l'hypothèse que les réacteurs sont indépendants et de loi de pannes identiques. Ces hypothèses ne sont pas réalistes du tout. D'ailleurs, les

[1] Voir notamment http://images.math.cnrs.fr/Accident-nucleaire-une-certitude.html et http://www.statistix.fr/spip.php?article87

[2] Le produit np donne l'espérance (moyenne théorique) du nombre d'accidents en n années. L'approximation de p' par np revient conceptuellement à dire que la probabilité d'avoir au moins deux accidents est négligeable, c'est en quelque sorte remplacer la loi binomiale par une loi de Bernoulli en 0 ,1 et pour cette loi, l'espérance est égale à la probabilité de 1.

450 réacteurs actuels n'ont pas tous 30 ans ! Nous sommes juste en train de faire des gammes en calcul de probabilités et non de jouer en concert. Le commentaire limpide de F. Sauvageot sur le site « image des maths » remet à ce propos quelques pendules à l'heure[3].

On suppose de plus qu'on peut dire qu'on a 30*450 réacteurs-années : l'intérêt de l'année est qu'on s'autorise à faire l'hypothèse supplémentaire que sur 1 an, le nombre de pannes pour chaque réacteur est au plus 1. Il est alors fondé de dire qu'on a un échantillon de taille $N=30\times 450$ d'une loi de Bernoulli de paramètre p, où p est petit (si la probabilité d'avoir une panne n'est pas petite, on ne pourrait pas négliger celle d'avoir deux pannes). La précision de l'estimation de p est « en $1/\sqrt{N}$».

Mais alors, si on passe en mois, il est tout aussi fondé de dire qu'on a un échantillon de taille $N'=12N$ de la probabilité p' d'avoir une panne en un mois, et comme on a plus de données, on aurait alors une meilleure précision (en $1/\sqrt{N'}$). Et si on passe en jours, ou en heures, etc., la taille du nombre des données observées ne cessera de grandir…et la précision de s'améliorer !

Cette multiplication spontanée du nombre d'observations, c'est tout simplement merveilleux !

Etudions très sommairement ce tour de passe-passe.

On estime p par R/N, la précision au niveau de confiance 95% s'écrit[4] $\delta/\sqrt{N}$. Soit :

$$R/N - \delta/\sqrt{N} < p < R/N - \delta/\sqrt{N}$$

On estime p' par R/N', la précision au niveau de confiance 95% s'écrit $\delta'/\sqrt{N'}$. Soit :

$$R/N' - \delta/\sqrt{N'} < p' < R/N' - \delta/\sqrt{N'}$$

La valeur de δ est environ $2\sqrt{p(1-p)} \approx 2\sqrt{p}$ (p est petit).

De même la valeur de δ' est environ $2\sqrt{p'(1-p')} \approx 2\sqrt{p'}$.

On a $N'=12N$ et comme on manipule des probabilités petites, on a :

$$p' \approx \frac{p}{12} \text{ d'où } \sqrt{\frac{p'}{N'}} = \frac{1}{12}\sqrt{\frac{p}{N}}.$$

Quand on dit que la précision de l'estimation augmente avec la taille de l'échantillon on sous-entend que le paramètre à estimer reste le même, alors qu'ici on augmente la taille de l'échantillon et on diminue en même temps la valeur du paramètre à estimer ! Le rapport entre la longueur de l'intervalle de confiance et le paramètre à estimer ne change pas quand on passe du mois à l'année, c.a.d que la

[3] Toujours à http://www.statistix.fr/IMG/pdf/proba_superieure_a_1.pdf

[4] On fait ici les calculs dans le cas où p tout en étant petit ne l'est pas trop, de telle sorte que l'approximation par une loi de Gauss soit possible (l'approximation n'a de toute façons pas de sens si la borne inférieure de l'intervalle de confiance ainsi trouvée est négative !). Nous avons considéré dans ce paragraphe des pannes et non des accidents majeurs graves justement pour pouvoir imaginer cette situation, l'approximation gaussienne conduisant à des formules simples pour la demi-longueur des intervalles de confiance.

précision relative (précision/paramètre) reste constante. Ou, pour le dire différemment, si on estime la probabilité d'avoir un accident en Europe en un an, la précision sera rigoureusement la même si on considère qu'on a 450 réacteurs sur 30 ans, 14 000 réacteurs-années, ou 12 ×14000 réacteurs -mois, ou 24× 12× 14000 réacteurs-heures !

3- La vraie valeur de *p ?*

Dans la discussion a propos de l'article dans Libération[5], Arnaud Lionnet propose l'exercice de statistique ci-dessous, où *p* est la probabilité pour un réacteur d'avoir un accident en un an et où « les journalistes » sont les auteurs de l'article de Libération (en fait, ils ne sont pas journalistes…)

Le premier exercice est d'estimer p. Le calcul des journalistes est juste. La valeur qu'ils obtiennent est un estimateur de p (estimateur étant le terme statistique pour ça). Ca veut dire qu'en principe la vraie valeur de p ne devrait pas être trop loin. Et vu qu'on demande à 14000 réacteur-an, on peut être assez confiant sur la valeur obtenue (dans un sondage typique, on demanderait à 1000 personnes s'ils vont voter pour le candidat A). Bien sûr on pourrait faire plus de statistiques et se demander combien pas-trop-loin de 0.0003 la vraie valeur de p est…

Comme il s'agit d'un exercice, on peut dire, employant un langage très « educnat » (éducation nationale) qu'il s'agit d'une situation *pseudo-concrète* et que partant de là, tout est permis. L'emploi du terme de « vraie valeur » de la probabilité *p* me semble néanmoins bien gênant. A. Lionnet compare la situation à un sondage, c'est-à-dire à un tirage de boules dans une urne contenant une proportion *p* de boules blanches : dans ce cas, si on dit que *p* est la *vraie* probabilité de tirer une boule blanche ça se comprend bien. Ici, on est quand même fort loin d'une telle situation, même métaphoriquement. Parler de la « *vraie valeur de p* » est un raccourci de langage (hélas ?) fréquemment utilisé entre statisticiens mais il ne favorise vraiment pas la compréhension de la statistique ! Il serait de même tout aussi peu pertinent de parler la *vraie valeur de la* probabilité pour le siècle qui vient :

-que la Manche gèle à Flamenville (d'où panne du circuit de refroidissement de la centrale située sur cette commune et accident)
-qu'un tsunami dévaste la France
-qu'une attaque terroriste ou un météorite détruise de nombreux réacteurs, etc.

Les probabilités dont on parle ici sont des paramètres de modèles (où issues de calculs dans des modèles dont on estime les paramètres). Un modèle est une reconstruction de la réalité dans le monde mathématique, et sa comparaison avec celle-ci ne se situe pas dans le registre du vrai ou du faux, mais du plus ou moins validé, du plus ou moins proche ou adéquat. L'adéquation de telles modélisations peut être cependant si remarquable (désintégration radioactive par un processus de Poisson par exemple) qu'on finit presque tous par se surprendre à penser que le modèle dit « le vrai », mais pour la probabilité d'accidents nucléaires, c'est autre chose.

[5] Même adresse http://www.statistix.fr/IMG/pdf/proba_superieure_a_1.pdf