L'ASSASSIN, LE JUGE ET LA STATISTIQUE

M. CHRISTINE [*] *et C. P. ROBERT* [**]

[*]INSEE - Unité "Méthodes Statistiques"
[**]Université Paris Dauphine - INSEE-CREST

Dans le cadre de l'instruction d'une affaire criminelle, la question suivante nous a été posée : peut-on, par un appel téléphonique passé sur un portable depuis la terrasse surplombant l'Etang de la Treille, à Teyron, activer la cellule 2802 de Nanteulles ? Une réponse négative conduirait en effet à mettre en contradiction le suspect avec les déclarations qu'il a faites sur son emploi du temps le 16 juillet 1969.

Pour tenter de répondre à cette question, une première expertise a éé réalisée par un expert des Télécommunications.

L'objet de cette étude a été de chercher à affiner les conclusions de la première expertise, en mettant en oeuvre différentes techniques statistiques et en confrontant leurs résultats.

1 Modélisation probabiliste de l'expérience

Le premier expert avait effectué une expérimentation statistique qu'il avait modélisée de façon à estimer la probabilité d'activation de cette cellule, soit p, et à construire un intervalle de confiance sur p.

L'expérimentation statistique qu'il a mise en oeuvre a consisté à faire, depuis l'Etang de la Treille, 77 essais d'appels téléphoniques successifs et mutuellement indépendants[1] et à observer le nombre de ceux qui auraient activé la cellule en cause. Ainsi, l'expert constate que ces essais sont tous négatifs, et en déduit un intervalle de confiance à 95% pour p, selon la théorie classique dans le cas des proportions avec approximation normale.

[1]Le caractère d'indépendance probabiliste des essais n'est pas formellement respecté dans l'expérience, en particulier du fait de la mémorisation des canaux par les mobiles.

Avenue Blaise Pascal
NANTEULLES
Borne 2802
25 km
Etang de la Treille
(Commune de TEYRON)
Autres
Bornes
TERRASSE

Cet intervalle (ici, unilatéral) permet d'affirmer que, à 95% de "chances", p est inférieur à à 3,39% (voir Section 2 et Annexe 1).

L'ensemble des 77 tests réalisés par l'expert est modélisé, en première approximation, par une distribution binomiale $\mathcal{B}(77, p)$, p étant la probabilité d'atteindre la borne 2802 de Nanteulles. *[Rappelons que la loi binomiale $\mathcal{B}(n, p)$ est celle de la somme d'une suite indépendante et identiquement distribuée (iid) de n essais qui conduisent à un succès avec probabilité p (voir Annexe 1)].*

Il s'agit bien d'une approximation car rien ne permet d'affirmer que tous les appareils utilisés ont la même probabilité d'atteindre cette borne, en particulier parce que certains ont été utilisés avec initialisation sur la borne 2802. Du point de vue de la modélisation probabiliste, il aurait été préférable d'utiliser uniquement le téléphone du suspect, ce qui ne semble pas avoir été le cas. De plus, il faut opérer sous l'hypothèse que les conditions lors du test sont identiques à celles supposées en vigueur le 16 juillet 1969. (L'expert a affirmé qu'il n'y avait effectivement pas de différence sur les deux zones entre la situation au 16 juillet 1969 et celle au moment des tests.)

Remarquons à ce stade que la réponse statistique liée à ces données ne pourra porter que sur la probabilité d'atteindre la borne de Nanteulles en téléphonant de l'étang de la Treille et, *en aucun cas*, sur la probabilité que le suspect ait passé un appel de l'étang de la Treille. En effet, les informations recueillies portent sur la probabilité de toucher la borne 2802 en étant à l'étang de la Treille et non sur la probabilité inverse d'être à l'étang de la Treille sachant que la borne 2802 a été touchée (ce qui, en termes probabilistes, correspond au théorème d'inversion dit *de Bayes*).

Pour déterminer cette probabilité inverse, il faudrait disposer d'une évaluation similaire des probabilités de toucher la borne 2802 pour l'ensemble des endroits où pouvait se trouver le suspect à 20h46 le 16 juillet 1969, ce qui ne peut malheureusement pas être mis en œuvre. (Se restreindre, comme unique alternative, à l'avenue Blaise Pascal[2] aurait pour effet de biaiser la réponse en défaveur de l'hypothèse "étang de la Treille" puisque la borne 2802 se trouve à proximité de cette avenue.)

[2]Domicile de la victime

2 Interprétations probabilistes des résultats

2.1 Introduction

Le fait que les 77 tests n'aient jamais atteint la borne 2802 peut être interprété de plusieurs façons.

Tout d'abord, on peut estimer la probabilité p d'activation de la borne 2802. L'estimateur habituel (estimateur du maximum de vraisemblance, sans biais optimal) donne ici: $\hat{p} = 0$.

On peut alors chercher à répondre directement à la question *"Est ce que p vaut 0 ?"*. Ceci correspond, en termes statistiques, à une problématique de test. Mais le test statistique est en fait impossible dans ce cas précis, à cause de la dégénérescence de la distribution binomiale pour $p = 0$. De plus tester l'exacte nullité de p n'a guère de sens, puisque le rapport d'expert indique que la possibilité de dépasser le seuil d'accès à la cellule 2802 ne peut être exclue.

2.2 Approximation normale

Une seconde approche statistique consiste à construire une région (ou intervalle) *de confiance* sur la probabilité p qui va s'écrire $[0, p_0]$ et vérifier la propriété

$$\text{Prob}(p \in [0, p_0]) = \alpha,$$

où α est choisi parmi les valeurs standards de $0,95$ ou $0,99$.[3] L'approximation normale fournit un intervalle dont la couverture est approximativement α (au sens où cette couverture tend vers α avec le nombre d'observations). Notons que, pour n résultats négatifs sur n essais, l'approximation normale donne pour borne supérieure

$$p_0 = \frac{u_\alpha^2}{u_\alpha^2 + n},$$

[3] Notons ici que la probabilité α associée à un intervalle de confiance $[0, p_0]$ ne signifie pas que p est supérieur à p_0 avec probabilité $1-\alpha$, mais plutôt que l'intervalle construit à partir des données, donc aléatoire, ne contient pas p dans $(1-\alpha)$ % des cas. Proposer l'intervalle de confiance $[0, 0.034]$ avec une confiance associée de 95% veut donc dire que l'intervalle $[0, 0.034]$ est une réalisation d'une variable aléatoire qui contient, avec probabilité $0,95$, la vraie valeur p et qui ne la contient pas avec probabilité $0,05$.

où u_α est le quantile de la loi normale standard au niveau α, soit $1,645$ au niveau $0,95$ et $2,33$ au niveau $0,99$ (voir Annexe 1). La Figure 1 donne l'évolution de p_0 en fonction de α. On trouve ainsi aux niveaux $0,95$, $0,99$ et $0,999$ les valeurs $0,034$, $0,066$ et $0,11$, respectivement.

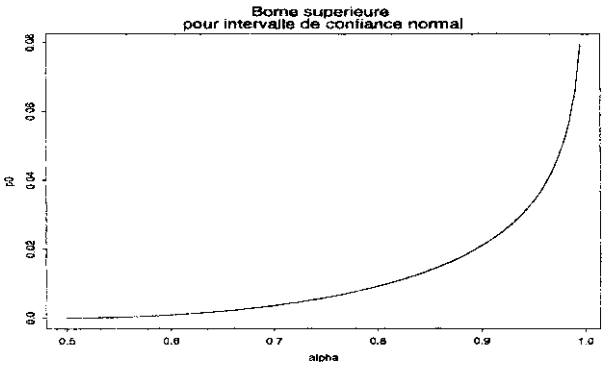

Figure 1: Evolution de la borne supérieure p_0 de l'intervalle de confiance normal en fonction de la confiance α.

2.3 Intervalle de confiance vrai : utilisation de l'inégalité de Cernov

L'inconvénient de l'approximation normale, pour utile, banalisée et acceptée qu'elle soit, est qu'on ne connaît pas en général (ou qu'on occulte) l'erreur commise en utilisant cette approximation.

Compte-tenu des enjeux attachés aux résultats de ces expériences, un calcul plus exact paraît préférable. Celui-ci peut être obtenu en construisant un intervalle de confiance par excès, via *l'inégalité de Cernov*

$$P_{p_0}(\hat{p}_n > p_1) \leq \inf_{s>0} \mathbb{E}\left[e^{s\{\hat{p}_n - p_1\}}\right] ,$$

où $\hat{p}_n$ représente la proportion observée des résultats positifs (dont l'occurrence sur une expérience élémentaire se produit avec une probabilité p_0). Dans

notre cas, il vient en fait que

$$P_{p_0}(p_0 > \hat{p}_n + a) \leq \left[\sup_{x \in [0,1-a]} m(x) \right]^n ,$$

où n est le nombre d'observations et

$$m(x) = \left(\frac{x}{x+a} \right)^{x+a} \left(\frac{1-x}{1-x-a} \right)^{1-x-a}$$

(voir annexe 2).

Cependant, la détermination explicite du nombre d'observations nécessaire (conduisant tous à un résultat négatif) ne peut se faire que par approximation numérique, les formules ne se résolvant pas. Pour les niveaux 0,95 et 0,99, on obtient ainsi les bornes 0,139 et 0,173, respectivement, c'est à dire que l'on peut affirmer, avec une probabilité de 0,95 (resp. 0,99) que p_0 est inférieur à 13,9% (resp. 17,3%).

2.4 Intervalle de confiance par inversion

Une autre approche possible est de produire un intervalle de confiance classique en inversant le test $H_0 : p = p_0$. Pour p_0 différent de 0 et de 1, la région d'acceptation de H_0 est de la forme $[p_1(p_0), p_2(p_0)]$. L'intervalle de confiance $[0, p_0]$ est alors fourni par la valeur limite de p_0 où l'observation 0 rejette le test, soit p_0 tel que $p_1(p_0) > 0$, qui vérifie

$$\text{Prob}_{p_0}(x = 0) = (1 - \alpha),$$

ce qui conduit à la borne

$$p_0(\alpha) = 1 - (1 - \alpha)^{1/n} .$$

La Figure 2 donne l'évolution de $p_0(\alpha)$ en fonction de α : on remarquera, par comparaison avec la Figure 1, que les valeurs de p_0 sont plus élevées pour les petites valeurs de α, mais que cette différence est minime. Ainsi, pour les niveaux 0,95, 0,99 et 0,999, on obtient pour $p_0(\alpha)$ les valeurs 0,038, 0,058 et 0,086, respectivement.

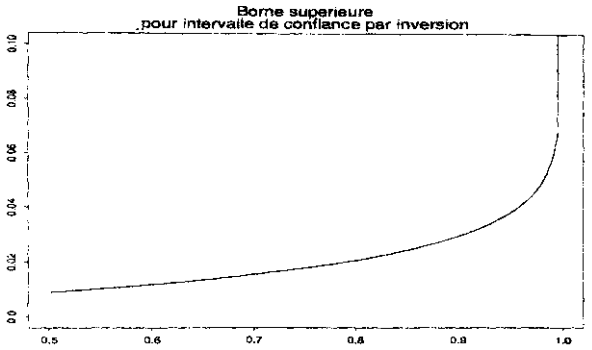

Figure 2: Evolution de la borne supérieure de l'intervalle de confiance par inversion en fonction de la confiance α.

2.5 Intervalle de confiance bayésien

Dans une perspective entièrement différente, dite *bayésienne*, on considère que la probabilité inconnue p suit une loi uniforme sur $[0, 1]$, de manière à ne privilégier aucune valeur ou zone de valeurs *a priori*. L'incorporation des 77 observations toutes négatives conduit alors à modifier cette loi sur p en une loi dite Beta (de première espèce) $\mathcal{B}e(78, 1)$, de densité $78\,p^{77}$. On peut alors calculer un intervalle de confiance sur p, soit $[0, p_0]$, en prenant le quantile au niveau α de cette loi. La Figure 3 fournit ces quantiles.[4] Les bornes aux niveaux 0,95, 0,99 et 0,999 sont alors 0,038, 0,058 et 0,086, en concordance totale avec l'approche par inversion.

2.6 Conclusion

Quelle que soit l'approche statistique choisie, *tout en restant dans le cadre d'une modélisation du problème où le paramètre p d'activation de la borne contient toute la problématique*, les mesures effectuées fournissent, aux seuils usuels, des intervalles de confiance qui ne permettent d'exclure catégorique-

[4]Notons que cette perspective bayésienne autorise la construction d'un test qui compare les lois marginales des observations sous les deux hypothèses. Dans notre cas, les densités marginales valent $m_0(0) = 1$ et $m_1(0) = \int_0^1 p^{77}\,dp = 1/78$, ce qui signifie que les données sont 78 fois plus en faveur de l'hypothèse nulle que de l'alternative.

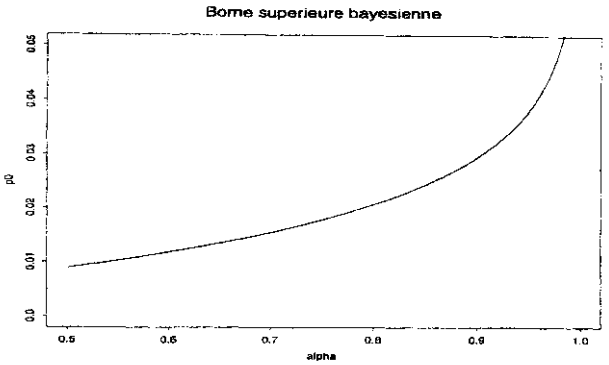

Figure 3: Evolution de la borne supérieure de l'intervalle de confiance bayésien en fonction de la confiance α.

ment la possibilité d'activation de la borne 2802 à partir de l'étang de la Treille.

3 Amélioration de la précision

L'avocat de la partie civile a considéré que la borne supérieure de cet intervalle de confiance était trop élevée et, surtout, que le niveau de confiance était trop faible pour que l'on puisse admettre, au vu des résultats expérimentaux, qu'il était *impossible* que le suspect ait pu téléphoner de l'endroit qu'il prétendait et activer ainsi la borne de Nanteulles.

Le seuil auquel l'avocat a fixé l'acceptation de l'affirmation ci-dessus est subjectivement établi à 0,22%, comme borne supérieure de l'intervalle de confiance pour p à 99%[5]. Dans ces conditions, il a préconisé que l'on procède à nouveau à des essais, en nombre approximatif de 3000 et dans des conditions de végétation analogues à celles de juillet 1969.

[5]Le "subjectif" est sans doute suggéré par le tableau de calculs du premier expert, donnant des intervalles de confiance pour p en fonction du nombre total d'essais lorsque la probabilité estimée est nulle.

Sous la contrainte de reproductibilité des expériences, à savoir la préservation de la probabilité p sous-jacente[6], il est indéniable que refaire une campagne de mesures avec un nombre d'essais supérieur (à 77) ne peut qu'améliorer la précision de l'estimation de p.

Notons ici que le chiffre de 3000 (en réalité, 2463) ne se justifie que parce qu'il permettrait, *si tous les essais étaient négatifs*, de parvenir à la conclusion souhaitée ci-dessus sur l'intervalle de confiance. Mais la conclusion serait différente si certains essais parmi les 3000 étaient positifs. La note du premier expert est d'ailleurs relativement claire sur ce point (*"le tableau suivant montre pour un nombre d'essais négatifs donné la probabilité maximum pour une fiabilité à 95% et 99%"*).[7] En revanche, l'avocat parle d'un nombre d'essais à effectuer sans tenir compte de l'éventualité d'essais à résultat positif.

Les analyses ci-dessus se transposent à un nombre quelconque d'essais. Les Figures 4 et 5 donnent les abaques correspondant aux cas normal et bayésien. (Le cas par inversion donne exactement la même courbe que la Figure 5.) Pour un seuil de fiabilité de 99%, en utilisant l'intervalle de confiance de Cernov, on trouve aussi que, pour pouvoir conclure que la probabilité d'activation est inférieure à $0,01$ (sachant qu'elle est inférieure à 0.05), ce nombre d'essais minimal est de 4088.[8] De même, l'approximation normale donne, pour le seuil de 99% et la borne $0,0022$, un nombre d'essais de 2463, et la solution par inversion donne 2091 essais, comme la solution bayésienne.

4 Autres modélisations

En l'état actuel des informations dont on disposait, il a été difficile de proposer et de tester une modélisation plus sophistiquée. La modélisation statistique proposée ici ramène le problème à l'estimation d'un unique paramètre,

[6] Condition impérative à respecter pour que les conclusions de cette nouvelle campagne puissent être comparées aux premières et être représentatives des conditions qui ont prévalu au moment des faits.

[7] Il serait encore plus précis de dire : *"pour un nombre donné d'essais conduisant tous à des résultats négatifs"*.

[8] Si on abaisse trop la valeur maximale de la probabilité, le nombre d'observations nécessaires sous l'approximation de Cernov croît très rapidement et n'a guère de sens pour $0,0022$.

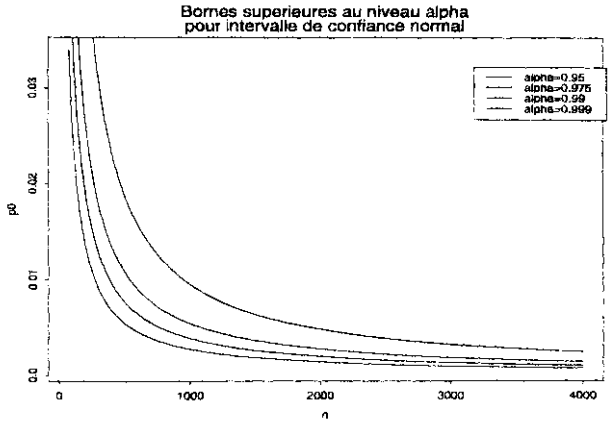

Figure 4: Evolution de la borne supérieure de l'intervalle de confiance normal en fonction de la taille de l'échantillon n, pour diverses valeurs de α.

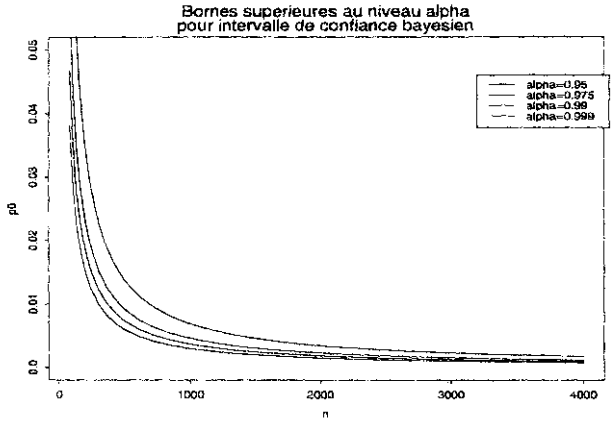

Figure 5: Evolution de la borne supérieure de l'intervalle de confiance bayésien en fonction de la taille de l'échantillon n, pour diverses valeurs de α.

p, censé résumer les caractéristiques aléatoires du phénomène étudié.

Pour progresser, il faudrait, idéalement, bien connaître les mécanismes d'activation d'une cellule à partir de l'émission d'un signal en provenance d'un téléphone mobile, comprendre pourquoi telle cellule est activée plutôt que telle autre et quels sont les facteurs susceptibles d'influencer ou de perturber cette activation. On peut imaginer en particulier que les localisations relatives du signal émis et de la cellule réceptrice jouent un rôle important, mais peut-être aussi la séquence des appels précédemment émis ou reçus.

Si une étude physique paraît trop complexe, compte-tenu du nombre de facteurs intervenant ou de l'impossibilité de les mesurer tous, une modélisation probabiliste du cheminement dans l'espace d'un signal à partir d'un point donné et/ou du niveau du champ électrique en un point donné en fonction du temps pourrait éventuellement porter ses fruits et ramener le problème à un petit nombre de paramètres à estimer.

Ceci reviendrait, dans une certaine mesure, à "expliquer" la probabilité p d'activation dont on a parlé ci-dessus à l'aide de facteurs explicatifs et d'hypothèses sur le comportement aléatoire de ces facteurs (distribution, nature des paramètres d'intérêt intervenant dans celle-ci etc..).

4.1 Tenir compte de la géométrie du phénomène

Une approche possible consisterait à dire que le champ théorique perçu en un point donné en provenance d'une cellule radio donnée est en moyenne une fonction de la distance entre les deux points considérés (le portable et la cellule) mais que les variations par rapport à ce niveau théorique dues à des perturbations extérieures ou des éléments supplémentaires non pris en compte dans cette fonction peuvent être modélisées par un aléa de type bruit blanc.

On aurait ainsi une relation de la forme

$$E = f(D) + U,$$

où E est la valeur du champ, D est la distance entre les deux points et U un aléa, dont on peut supposer qu'il suit la loi normale $\mathcal{N}(0, \sigma^2)$, tout en étant

indépendant de D.

La distance D dépend de la localisation de la cellule radio, parfaitement connue et de celle du portable appelant, inconnue, mais sur laquelle on peut faire des hypothèses raisonnables, compte tenu du fait que l'on cherche à savoir si l'appel a eu lieu d'une zone relativement circonscrite (une loi uniforme sur un cercle ou un rectangle voisinant le lieu d'appel supposé pourrait convenir).

Suivant l'information que la cellule n'est susceptible d'être activée que si le niveau du champ reçu est supérieur à un certain seuil, on pourrait, moyennant ces différentes hypothèses, évaluer la probabilité pour que le champ soit supérieur au seuil d'accès (supposé connu et ne dépendant que de la cellule radio).

Cependant, cette approche se heurte à plusieurs difficultés :

1. il faut estimer les paramètres intervenant dans la modélisation (essentiellement, le paramètre σ intervenant dans les aléas qui figurent dans l'expression du champ) et, pour ce, mener une campagne de mesures pour chacune des antennes susceptibles d'intervenir.

2. il faut faire des hypothèses sur la loi de déplacement du portable dans la zone considérée.

3. il n'est pas certain qu'on puisse construire explicitement la fonction $f(D)$ indiquée ci-dessus, ni même que la valeur du champ ne dépende que de cette distance : ainsi, pour une même valeur de D, il est possible qu'on trouve différentes valeurs du champ car l'inclinaison du portable, l'orientation relative de l'antenne de celui-ci et de la cellule radio, le dénivelé entre les deux points, la plus ou moins grande couverture végétale etc..., peuvent jouer.

4. la connaissance du champ émis par les différentes antennes ne suffit pas à déterminer laquelle est activée par le portable, puisque ce n'est pas la loi du plus fort champ qui l'emporte.

4.2 Analyse normale des niveaux de champ

Une manière plus simple de modéliser la situation, en s'appuyant sur les données disponibles, consisterait à combiner dans un même aléa l'incertitude due à la position du portable et celle liée aux causes non maîtrisées et non mesurées de variation du champ électrique.

Ainsi, on serait amené à considérer que la valeur du champ en provenance d'une cellule i dans la zone considérée (sans davantage de précision) est un aléa, par exemple normal, de paramètres inconnus μ_i et σ_i^2. Les différentes mesures de champ effectuées lors de la première expertise permettent d'estimer ces paramètres, à condition de disposer des données numériques exactes et d'être assuré de leur caractère i.i.d. On peut alors en inférer la probabilité pour que, de la zone considérée, le niveau de champ de la cellule 2802 soit supérieur au seuil d'accès de -101dBm.

On pourrait, plus généralement, s'appuyer sur la cartographie des champs réalisée par le premier expert, afin de donner une estimation de la valeur moyenne du champ sur un maillage du territoire et d'estimer ainsi, pour toute maille élémentaire, la probabilité qu'un téléphone placé dans cette maille reçoive des cellules considérées un champ supérieur à leur seuil d'accès. Toutefois, l'étape suivante consistant à évaluer la probabilité pour que, en un point donné, ce soit la cellule 5479, plutôt que la cellule 2802, qui soit activée, est délicate à mettre en œuvre, car elle dépend de nombreux facteurs, et pas seulement de la différence des champs reçus en provenance de ces cellules.

Sans aller jusqu'à cette étape, il peut cependant y avoir intérêt à mobiliser les données de mesure du champ sur la zone considérée, en ajoutant éventuellement de nouvelles observations aux données existantes. On peut penser en effet que des mesures de champs supplémentaires pourraient avoir une valeur probante, même en cas de modification du réseau[9], ce qui n'est pas nécessairement le cas de mesures d'activation comme celles rappelées en §1.

En tout état de cause, si des observations complémentaires devaient être menées, il serait essentiel d'assurer leur caractère indépendant et équidistribué,

[9]Si l'on admet que la valeur du champ reçu en un point ne dépend pas fortement de l'environnement radio-électrique.

afin qu'une modélisation simple (échantillon normal) puisse être utilisée.

ANNEXE 1

Intervalle de confiance normal pour une proportion lorsque l'estimation (naturelle) de celle-ci est nulle.

Soit $\{X_i\}_{i\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes suivant toutes la même loi de BERNOULLI $\mathcal{B}(1,p)$. p représente la probabilité qu'une variable X_i donnée prenne la valeur 1.

Si l'on fait n tirages des variables X_i, la proportion empirique $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ s'interprète comme la fréquence d'apparition de la valeur 1. C'est un estimateur sans biais de la vraie probabilité p. Sa variance vaut: $V\hat{p}_n = \frac{p(1-p)}{n}$. Quand n tend vers l'infini, le théorème central limite indique que:

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{Loi} N(0, p(1-p)) \ .$$

Si l'on remplace la loi exacte[10] (mais relativement compliquée à écrire) de $\sqrt{n}(\hat{p}_n - p)$ par sa loi limite, on peut alors construire un intervalle unilatéral de confiance asymptotique pour p en écrivant:

$$P\left[\frac{\sqrt{n}(p-\hat{p}_n)}{\sqrt{p(1-p)}} \leq K\right] = \alpha$$

où, α étant donné, K est déterminé comme quantile d'ordre α de la loi normale $N(0,1)$.

La résolution mathématique rigoureuse de l'inégalité $\sqrt{n}\frac{(p-\hat{p}_n)}{\sqrt{p(1-p)}} \leq K$ en p conduit à une double inégalité qui se simplifie quand $\hat{p}_n = 0$.

Dans ce cas, on a en effet:

$$\sqrt{n}\frac{p}{\sqrt{p(1-p)}} \leq K$$

soit: $\frac{np^2}{p(1-p)} \leq K^2$ ou encore: $\underline{p \leq \frac{K^2}{n+K^2}}$.

[10] que l'on peut obtenir en remarquant que $n\hat{p}_n \sim \mathcal{B}(n,p)$

Ainsi, avec $n = 77$ et $\alpha = 95\%$, on trouve $K = 1,64$, d'où $\underline{p \leq 3,39\%}$.

Si l'on veut que, pour $\alpha = 99\%$, la borne supérieure de p soit inférieure à $0,22\%$, cela donnera la valeur de n minimale,

soit $\underline{n \geq K^2 \left(\frac{1}{0,0022} - 1 \right)}$.

Ici, K vaut $2,33$, d'où: $\underline{n \geq 2463}$.

ANNEXE 2

Déviation d'une moyenne empirique par rapport à son espérance (au moyen de l'inégalité de CERNOV).

On reprend les hypothèses et les notations de l'Annexe 1.

Pour a positif donné on a:

$$\begin{aligned}
P[\hat{p}_n > p + a] &= P[e^{s\hat{p}_n} > e^{s(p+a)}] && \forall\, s > 0 \\
&\leq e^{-s(p+a)} E e^{s\hat{p}_n} && \forall\, s > 0 \quad \text{[inégalité de MARKOV]} \\
&= e^{-s(p+a)} E e^{\frac{s}{n}\sum_{i=1}^{n} X_i} && \forall\, s > 0 \\
&= e^{-s(p+a)} \prod_{i=1}^{n} E e^{\frac{s}{n} X_i} && \forall\, s > 0 \quad \text{[par indépendance des } X_i\text{]} \\
&= \left[e^{-\frac{s}{n}(p+a)} L_X\left(\frac{s}{n}\right) \right]^n
\end{aligned}$$

où $L_X(u)$ représente la transformée de LAPLACE commune des variables X_i :

$$L_X(u) = E e^{uX_i} .$$

On en déduit alors l'inégalité de CERNOV:

$$P[\hat{p}_n > p + a] \leq \left[\min_{u>0} e^{-u(p+a)} L_X(u) \right]^n .$$

Dans le cas présent:

$$L_X(u) = pe^u + 1 - p .$$

La fonction $u \to h(u) = e^{-u(p+a)} L_X(u) = e^{-u(p+a)}(pe^u + 1 - p)$
possède un minimum en un point annulant sa dérivée ou, mieux, sa dérivée logarithmique,
soit: $\dfrac{h'(u)}{h(u)} = -(p+a) + \dfrac{pe^u}{pe^u + 1 - p}$.

La solution u^* vérifie donc, après calculs: $e^{u^*} = \dfrac{(p+a)(1-p)}{(1-p-a)}$)
et le minimum de h vaut:

$$h^* = \left(\frac{p}{p+a}\right)^{p+a} \left(\frac{1-p}{1-p-a}\right)^{1-p-a} .$$

Notons $m(p)$ cette quantité.

On en déduit que:

$$P[\hat{p}_n > p + a] \leq [m(p)]^n .$$

On aura de même, en changeant p en $1 - p$:

$$P[1 - \hat{p}_n > 1 - p + a] \leq [m(1-p)]^n .$$

Soit :

$$P[p > \hat{p}_n + a] \leq \left[\left(\frac{1-p}{1-p+a} \right)^{1-p+a} \left(\frac{p}{p-a} \right)^{p-a} \right]^n \forall\, p \in]0,1[.$$

Au moyen de cette dernière inégalité, on peut construire un intervalle de confiance vrai, et non plus approché pour p.
En effet, si l'on s'arrange pour rendre la probabilité $P[p > \hat{p}_n + a]$ inférieure à un seuil donné $1 - \alpha$, alors on aura: $P(p \leq \hat{p}_n + a) \geq \alpha$ et l'on obtiendra un intervalle de confiance pour p de la forme $]0, \hat{p}_n + a]$ ayant une probabilité au moins égale à α.

La condition $P[p > \hat{p}_n + a] \leq 1 - \alpha$ est obtenue dès que: $[m(1-p)]^n \leq 1 - \alpha$.
L'inconvénient est que cette condition fait apparaître explicitement le paramètre p inconnu.
On peut néanmoins s'en sortir en écrivant:

$$\forall\, p \in]0,1[: P[p > \hat{p}_n + a] \leq [m(1-p)]^n$$

donc, a fortiori:

$$\forall\, p \in]a,1[: P[p > \hat{p}_n + a] \leq [m(1-p)]^n \leq \left[\sup_{x \in]a,1[} m(1-x) \right]^n .$$

On sait que: $\forall\, p \in]0,1[: m(1-p) \in]0,1[$. En effet, $m(1-p)$ est le minimum, sur $\mathbb{R}^{+*}$, de la fonction L_{1-x} définie par $L_{1-x}(u) = (1-p)e^u + p$, qui vaut 1 en 0, a une pente négative $(1-p)$ en 0 et tend vers $+\infty$ en $+\infty$, donc admet un *minimum strictement inférieur à 1 et évidemment strictement positif.* Il faut alors s'assurer que $\sup_{x \in]a,1[} m(1-x)$ n'est pas égal à 1 pour que l'inégalité ci-dessus soit utilisable.

Or, si l'on étudie la fonction $x \to m(1-x) = \left(\frac{1-x}{1-x+a} \right)^{1-x+a} \left(\frac{x}{x-a} \right)^{x-a}$, on constate aisément quelle admet pour limites: 0 en 1 et $1 - a$ en a. On peut donc la prolonger par continuité sur $[a,1[$ et, sur cet intervalle, elle atteint son maximum qui a donc une valeur strictement inférieure à 1, soit $q^*(a)$.

On notera que ce maximum n'est pas atteint en a, mais en un point supérieur, donc qu'il a une valeur strictement supérieure à $1-a$.
(Cf. Annexe 3)

Au total, on a:

$$\forall\, p \in]a, 1[\colon P[p > \hat{p}_n + a] \leq [q^*(a)]^n \ .$$

Les observations empiriques ayant conduit à $\hat{p}_n = 0$, l'intervalle de confiance pour p sera de la forme $]0, a]$. Le seuil a étant fixé $(0,0022$ par exemple), on en déduit (au moins de manière numérique, à défaut de pouvoir la calculer explicitement) la quantité $q^*(a)$ et on choisit n tel que $[q^*(a)]^n \leq 1-\alpha$, où α est le seuil de fiabilité (99% par exemple).

On trouve, numériquement, que le maximum est atteint en la valeur $0,50073$ (à 5×10^{-6} près) et qu'il vaut $0,99999032$.

Au seuil de fiabilité choisi, on obtient alors: $\underline{n \geq 475741}$.

Si on admet, au regard des expériences réalisées que la probabilité p ne peut excéder 0,05, on gagne en précision en remplaçant $q^*(a)$ par $m(1-0,05) = m(0,95)$. On trouve alors, numériquement, $n \geq 4088$ au seuil de 99% et 2659 au seuil de 95%.

ANNEXE 3

Etude de la fonction $x \to m(1-x) = \left(\frac{1-x}{1-x+a}\right)^{1-x+a} \left(\frac{x}{x-a}\right)^{x-a}$ sur $]a, 1[$

Posons: $g(x) = \log m(1-x)$.

On a, après calculs:

$$\begin{aligned} g(x) &= (1-x+a)\log\frac{1-x}{1-x+a} + (x-a)\log\frac{x}{x-a} \\ g'(x) &= \log\frac{x(1-x+a)}{(x-a)(1-x)} - \frac{a}{x(1-x)} \\ g''(x) &= \frac{-1}{(1-x+a)(x-a)} + \frac{1}{x(1-x)} + a\frac{1-2x}{x^2(1-x)^2} = \\ &\quad \frac{a}{x(1-x)}\left[\frac{2x-1-a}{(1-x+a)(x-a)} + \frac{1-2x}{x(1-x)}\right] \end{aligned}$$

Etudions le signe de $g''(x)$.

$$\begin{aligned} g''(x) > 0 \quad &\Leftrightarrow \quad \frac{2x-1-a}{(1-x+a)(x-a)} > \frac{2x-1}{x(1-x)} \\ &\Leftrightarrow \quad (2x-1)x(1-x) - ax(1-x) > (2x-1) \\ &\qquad [x(1-x) + ax - a(1-x) - a^2] \\ &\Leftrightarrow \quad -x(1-x) > (2x-1)(2x-1-a) \\ &\Leftrightarrow \quad 3x^2 - x(3+2a) + 1 + a < 0 \end{aligned}$$

Or le trinôme $x \to 3x^2 - x(3+2a) + 1 + a$ a pour discriminant:

$$\nabla = (3+2a)^2 - 4.3(1+a) = 4a^2 - 3\ .$$

Ce discriminant est positif si et seulement si $a^2 \geq \frac{3}{4}$. Comme, dans le contexte étudié, a est très petit, *cette condition n'est pas réalisée.*

On en déduit que le trinôme considéré est sans racine et *reste toujours positif.* La condition $g''(x) > 0$ n'est donc jamais réalisée.

On en déduit que: $\forall\ x \in]a, 1[$: $g''(x) < 0$, donc g' *est décroissante.*

Or, si l'on étudie les limites de g' aux bornes de $]a, 1[$, on obtient aisément:

$$\begin{aligned} \lim_{x \to a} g'(x) &= +\infty \\ \lim_{x \to 1} g'(x) &= -\infty \end{aligned}$$

g' s'annule donc 1 fois et 1 seule sur $]a, 1[$.

On en déduit le tableau de variation suivant:

x	a		1
g	$-\infty$	$\searrow$	$+\infty$
		0	
g	$\nearrow$		$\searrow$
$m(1-x)$		$q^*(a)$	
	$1-a$	$\nearrow$	0 $\searrow$