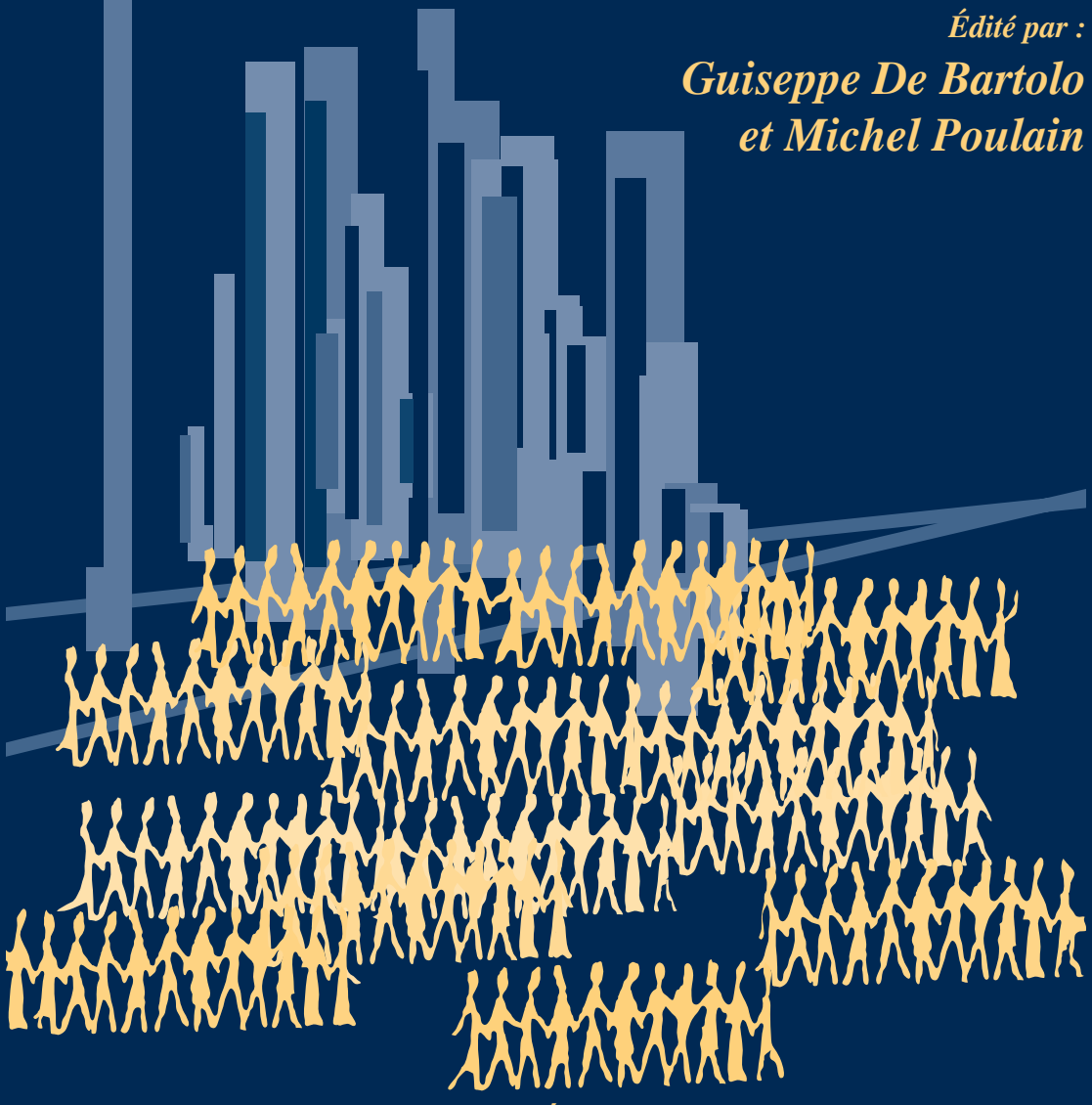
La démographie appliquée à la gestion publique et des entreprises

Actes du séminaire de l'AIDELF en Calabre (Cosenza, avril 1995), et de la session spéciale du Congrès de l'EAPS (Cracovie, juin 1997)

Édité par :
Guiseppe De Bartolo et Michel Poulain



ASSOCIATION INTERNATIONALE DES DÉMOGRAPHES DE LANGUE FRANÇAISE
AIDELF

Comment appréhender le problème statistique des petits nombres en démographie ?

Pierre ARS[1]
Institut de Statistique - Université catholique de Louvain, Belgique

Luc DAL[2] **et Michel POULAIN**[3]
GéDAP - Université catholique de Louvain, Belgique

Dans cet article, nous discutons des problèmes statistiques liés à l'analyse démographique de petites populations. Dans ce but, nous construisons des intervalles de confiance pour des taux, quotients ou proportions. Dans le cas de (petites) populations, il est possible de construire des intervalles exacts, tandis que dans les autres cas, on détermine des intervalles de confiance plus précis que ceux utilisés d'ordinaire. En outre, on propose un intervalle de confiance approché et un test exact pour la différence de deux proportions. Les différentes méthodes sont appliquées à deux situations réelles.

1. Les objectifs poursuivis

En démographie, on considère principalement deux types de variables à partir desquelles on peut étudier la population relative à un territoire donné :

1. l'effectif de cette population et de ses différentes sous-populations (selon l'âge, le sexe, l'état matrimonial, la nationalité, ...). Ces nombres sont des entiers non négatifs correspondant à des observations à un instant donné « t ». En les rapportant les uns aux autres, on calculera généralement des proportions caractérisant différentes sous-populations.
2. des nombres d'événements (naissances, décès, mariages ou migrations) rendant compte du mouvement de la population pendant une période donnée [t, t+1]. Sur base de ce second type de variables, on définit le plus souvent des taux en rapportant le nombre d'événements à la population moyenne soumise au risque.

Qu'en est-il de l'importance du caractère aléatoire de ces mesures ou dénombrements?

Tout d'abord, on ne peut exclure que, dans certains cas, la mesure démographique en elle-même (le dénombrement des structures ou la comptabilité des événements) soit également entachée d'une marge d'erreur. Cette marge d'erreur pourra comprendre une erreur de type systématique de sous-dénombrement, par exemple, mais également une erreur de type aléatoire qui trouve son origine au cours des différentes phases de la méthode de la collecte. Ce phénomène se constate très bien dans le dénombrement des migrations internes à un pays donné qui font l'objet d'une double mesure de façon quasi indépendante. En fait, tout courant de migration interne est comptabilisé avec un même outil statistique au titre d'émigrations par le lieu de départ et au titre d'immigrations par le lieu de destination. Pratiquement, on observe que les chiffres relatifs à un même courant de migration diffèrent et que la différence entre les deux mesures est proportionnelle à la racine carrée du volume du courant de migration que l'on

[1] ars@stat.ucl.ac.be
[2] dal@spri.ucl.ac.be
[3] poulain@spri.ucl.ac.be

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

cherche à mesurer. Il diminue par conséquent en termes relatifs avec la racine carrée de l'importance de ces courants de migration[4].

Par ailleurs, nous nous placerons résolument dans le cadre de l'hypothèse avancée par Keyfitz en 1966 : rappelons que celle-ci postule que la population étudiée n'est autre qu'un échantillon extrait d'une population hypothétique de taille infinie. Prenons un exemple : si la population infinie comprend un certain pourcentage (inconnu) de personnes âgées de 60 ans et plus, la population finie de taille n étudiée affichera, quant à elle, une proportion qui ne sera qu'une estimation de ce pourcentage. Plus la taille n de la population sera importante et plus l'estimation sera meilleure et assortie d'un intervalle de confiance plus réduit.

Enfin, dans le cas spécifique du dénombrement des événements au cours d'une période donnée en matière de fécondité, mortalité ou mobilité spatiale, on est évidemment dans l'impossibilité de maîtriser toutes les variables (comportementales, physiologiques, socio-économiques et autres) pouvant rendre compte de l'occurrence de ces événements. On préférera par conséquent admettre qu'il s'agit d'une occurrence à caractère aléatoire et que le nombre d'événements effectivement observés entre deux instants t_0 et t_1 sera le résultat d'un processus qui comprend une composante aléatoire.

Finalement, peu importe si le caractère aléatoire des mesures démographiques peut se justifier par l'une ou l'autre, voire plusieurs, de ces hypothèses. Ce qui prime, c'est d'estimer l'ampleur de ce caractère aléatoire au niveau de l'analyse et de l'interprétation des indices démographiques que nous calculons couramment. C'est principalement lorsque les populations sont de taille réduite, et/ou que les événements considérés sont rares que ce caractère aléatoire se traduit par un risque d'interprétation erronée des indices calculés. Force est donc de constater que toutes les mesures démographiques sont affectées d'une composante aléatoire qui nécessite le recours à un intervalle de confiance, celui-ci étant d'autant plus restreint que la taille de la population concernée est grande. Dès lors, en matière de petites populations et de petits nombres, les deux questions principales qui se posent au démographe sont les suivantes :

1. indépendamment de toute erreur de mesure, une proportion ou un taux observés sur une population comprend toujours une marge d'erreur (puisque cette proportion est une estimation de la vraie proportion qui elle est inconnue). On constate empiriquement que cette marge d'erreur est, entre autres, une fonction décroissante de la taille de cette population. De quelle manière varie cette incertitude, en fonction de la taille de la population, et quelle confiance peut-on accorder à une telle proportion ou taux ? Autrement dit, quelle précision ou quel crédit peut-on donner à une telle mesure ? La réponse à cette question est évidemment essentielle dans le cadre de l'analyse que l'on fera à partir de cette observation.
2. corrélativement, les différences entre les valeurs des indices démographiques calculés pour deux populations de tailles différentes reflètent-elles des situations ou des comportements résolument distincts, ou sont-elles plutôt le fait d'un processus stochastique lié à la petite taille des populations soumises au risque et/ou aux faibles taux d'occurrence des événements démographiques, le tout se traduisant par un nombre limité d'individus ou d'événements observés ? En d'autres termes, la valeur de cet indice pour la première population est-elle significativement supérieure ou inférieure à celle calculée pour la seconde population ?

On se référera éventuellement à l'annexe 1 pour un rappel succinct des principales distributions statistiques utilisées en démographie.

[4] Des considérations théoriques permettent d'expliquer cette constatation.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

2. Applicabilité des distributions statistiques discrètes usuelles aux phénomènes démographiques

Au sein d'une population de taille n, deux types de variables sont donc prises en considération :

– des événements qui peuvent se réaliser avec une probabilité θ et qui modifient les structures ;

– des structures qui partitionnent la population (selon l'âge, le sexe, l'état civil , …).

Lorsqu'on étudie cette population, on peut dès lors se poser deux types de questions :

– le premier porte sur les événements : on peut, par exemple, se demander quelle est la probabilité qu'un nombre X d'individus de la population vivent un type d'événement.

– le second type de questions concerne les structures : on peut se demander quelle est la probabilité que Y individus (choisis au hasard au sein de la population supposée parfaitement connue) appartiennent à une structure donnée S dans la population.

Le premier type de questions est un problème de nature statistique étant donné qu'on ignore les paramètres qui définissent la loi de probabilité d'occurrence des événements étudiés et qu'il faut tout d'abord estimer ces paramètres (ou les tester). Le second type de questions est un problème de probabilité, puisqu'il n'y a aucun problème d'estimation. Cet article ne considère que le premier type de problème.

2.1. La distribution hypergéométrique

Celle-ci modélise le problème de l'extraction (non exhaustif) de n individus d'une population de taille N dont R individus ont une caractéristique donnée et N-R ont la caractéristique opposée.

Si on traite un problème de sondage ou d'échantillonnage au sein d'une population finie, il peut être justifié de faire usage de la distribution hypergéométrique. En prenant un modèle d'extraction d'urne, si on connaît R (ou N-R), il n'y a aucun problème d'estimation et on est en mesure de calculer immédiatement la probabilité recherchée (en recourant éventuellement à l'approximation binomiale si la taille de la population est grande). Si par contre on ne connaît pas exactement R (ou N-R), alors on se trouve face à un problème d'estimation à partir d'un échantillon extrait d'une population de taille finie et ce problème se situe hors de notre contexte, puisque nous avons retenu l'hypothèse de Keyfitz.

Par contre, si on s'intéresse à l'occurrence d'événements, on se tournera vers la loi binomiale dont on essayera d'estimer le paramètre : même dans le cas de petites populations, la distribution hypergéométrique n'est pas appropriée pour traiter ce type de problème, puisque, avant leur réalisation, on ignore combien d'événements vont se produire : en reprenant le modèle d'urne, on ne connaît exactement pas R, le seul paramètre connu est N, et on se trouve face à un problème d'estimation.

Dès lors, dans notre problématique, l'usage de la distribution hypergéométrique (exacte ou approchée par une distribution binomiale de paramètres connus) constituerait une erreur méthodologique.

2.2. La distribution binomiale

Celle-ci repose sur deux hypothèses (indépendance des événements[5] et constance de la probabilité) que nous examinons dans le contexte des phénomènes démographiques.

---

[5] L'indépendance de deux événements A et B s'exprime par $P[A] = P[B] = P[A \cap B]$ ou encore $P[A|B] = P[A]$ lorsque $P[B] \neq 0$. Cela signifie que la réalisation de B n'a aucun effet sur la réalisation de A.

Première hypothèse : au cours d'une période de temps donnée, on suppose que tous les individus d'une population sont soumis à un risque qui se caractérise par l'occurrence d'un type d'événement. Chaque individu peut vivre cet événement et ceci indépendamment des autres individus. Par exemple, au cours de la période de temps donnée, des individus vont décéder et ces décès auront lieu indépendamment les uns des autres : a priori, il n'y a pas de raison de penser que le décès d'une personne A ait une influence quelconque sur le décès d'une personne B.

On pourrait objecter que dans certaines situations, telles que en présence d'épidémies, de catastrophes, … il y a un nombre plus important d'événements et donc que le comportement d'un individu A a une influence sur celui d'un individu B et donc que l'on perd l'indépendance. En réalité, il n'en n'est rien, mais dans ces cas, c'est le niveau du risque qui est beaucoup plus élevé. Par exemple, rien ne permet d'affirmer que si un individu A décède cela aura une influence sur la propension d'un autre individu B à mourir. Ce raisonnement peut être généralisé aux autres risques démographiques (par exemple la natalité) et il est donc justifié d'accepter l'hypothèse d'indépendance.

Seconde hypothèse : d'emblée, cette hypothèse pourrait être mise en doute : par exemple dans le cas de la mortalité, on peut objecter que tous les individus ne sont pas affectés d'une probabilité de décès identique, celle-ci dépendant d'un ensemble de facteurs individuels (sexe, âge, état matrimonial, profession, …) et comportementaux. Cependant, en l'absence d'informations détaillées, on est forcé d'admettre que le niveau du risque étudié est le même pour toute la population soumise à ce risque. Seule une analyse plus détaillée, par exemple de la mortalité selon l'âge, le sexe, … permettrait de mettre en évidence ces différences, mais elle se situerait dans le cadre de l'analyse explicative et requerrait de ce fait des informations plus détaillées.

Notons cependant que si on stratifie une population en classes homogènes pour le risque considéré (ce qui est théoriquement possible si on travaille, par exemple, pour la mortalité) alors, au sein d'une même strate, la probabilité de subir le risque est constante, par définition. Dans ce cas, un autre problème se pose et qui est lié à un problème d'ajustement. Supposons que l'on ait stratifié la population en k classes caractérisées par les probabilités (θ_1, …, θ_k). Au sein d'une strate i, le nombre d'événements observés X_i est une variable aléatoire binomiale $Bi(n_i,\theta_i)$ où n_i est la taille de la population de la strate i. Le problème qui se pose alors est de trouver la distribution de la somme des X_i, car la somme de variables binomiales n'est pas, en général, une binomiale. Toutefois, nous avons constaté sur des données réelles que X suit approximativement une loi binomiale $Bi(n,\theta)$ où n est la somme des n_i et θ est la moyenne des θ_i pondérée par les n_i. Ceci s'explique par la règle de Bayes en considérant un mélange de k populations, chacune ayant un poids . Dans ce cas, la probabilité qu'un individu choisi au hasard subisse le risque considéré est alors θ si on admet l'hypothèse d'indépendance ; on retombe alors sur les conditions conduisant à une variable binomiale. Néanmoins, X n'est pas une binomiale de paramètres n, θ et cet argument ne fait que de justifier intuitivement la qualité de l'ajustement.

2.3. La distribution de Poisson

Bien qu'elle soit plus restrictive que la distribution binomiale (puisque 3 hypothèses sont formulées au lieu de 2), la distribution de Poisson peut être utile pour modéliser l'occurrence des événements démographiques, à condition de se placer sur l'axe du temps[6]. Par ailleurs, on rappelle à l'annexe 2 que le loi de Poisson est un cas limite de la loi binomiale (« lorsque n devient grand et θ est petit », ce qui est souvent le cas en démographie).

Première hypothèse : à condition de découper l'intervalle de temps de manière suffisamment fine, il est légitime de penser que deux phénomènes démographiques ou plus ne

[6] Dans le cadre d'une analyse spatiale, on pourra faire usage de la distribution de Poisson et se placer dans le plan, les événements étant alors des points.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

peuvent survenir exactement au même instant (sauf peut-être dans le cas de la nuptialité où nécessairement deux individus se marient au même instant : mais dans ce cas, il n'y a qu'un seul mariage !). Cette hypothèse est donc admissible.

Deuxième hypothèse : celle-ci est assez naturelle : en se restreignant au terme du premier ordre, on peut admettre que le nombre d'événements qui se produisent pendant un intervalle de temps assez court est directement proportionnel à la longueur de cet intervalle de temps.

Troisième hypothèse : l'hypothèse d'indépendance a été discutée et admise au point précédent.

Chacune de ces distributions présente des avantages et des inconvénients : la loi de Poisson est plus facile à manipuler, elle est additive et sa moyenne coïncide avec son espérance, mais elle est plus restrictive. La loi binomiale est plus lourde au niveau des calculs et n'est pas additive, mais en contrepartie, elle est moins restrictive et semble (légèrement) mieux répondre à la réalité.

Aussi, par souci de généralité, le caractère binomial du nombre des événements est celui qui sera retenu dans les lignes qui suivent.

3. Construction des intervalles de confiance et tests

Dans la problématique qui est la nôtre, il s'agit donc :

1. de déterminer des intervalles de confiance pour une proportion θ_0 observée et à partir de là, de calculer des intervalles de confiance sur le nombre d'événements qui définissent θ_0 ;
2. de mettre en évidence, ou de tester, la nullité, ou la non nullité, de la différence entre deux proportions.

Ces deux points font l'objet des paragraphes 3.2 et 3.3.

Avant de les développer, il est essentiel de faire un bref rappel et de repréciser les notions de test et d'intervalles de confiance. La confusion qui règne entre les deux concepts est à l'origine d'erreurs fréquentes. L'élaboration des tests et la recherche d'intervalles de confiance sont deux problèmes relativement proches, mais toutefois différents et ils répondent à des problèmes statistiques distincts.

3.1. Préliminaires

3.1.1. Intervalles de confiance

La construction d'un intervalle de confiance répond à un problème d'estimation : on observe une valeur d'une variable aléatoire dont la distribution dépend d'un paramètre[7], on se fixe une probabilité élevée (appelée le niveau de confiance qui est fixé souvent à 90 ou 95 %), et à partir de là, on détermine une région qui contient le paramètre observé avec cette probabilité. Si le paramètre est réel, on peut construire des intervalles bilatères[8] ou unilatères. Les premiers sont de la forme]T ; T'[, tandis que les seconds sont de la forme [0 ; T[ou]T' ;

[7] Celui-ci peut être réel ou vectoriel.

[8] A ce propos, signalons que l'on trouve dans la littérature deux définitions différentes pour les intervalles de confiance bilatères pour un paramètre θ au niveau $1-\alpha$. On trouve, selon les auteurs, soit $P[T_1 < p < T_2] = 1-\alpha$ (i), soit $P[T_1 \leq p \leq T_2] = 1-\alpha$ (ii). Par exemple, Bickel et al. ainsi que Dagnelie utilisent (ii), tandis que Mood et al. ainsi que Saporta utilisent (i). Sans le mentionner explicitement, Bickel et al. utilisent toutefois la définition (i) lorsqu'ils se placent dans le cas discret. Cette différence (minime) entre les deux définitions n'a pas d'importance dans le cas de distributions continues, mais, par contre, elle en a une dans le cas discret. La distinction est souvent passée sous silence du fait que l'on travaille essentiellement avec des variables continues. Nous retiendrons pour notre part la définition (i) et nous ne considérerons dans l'exposé théorique que les intervalles de type bilatères.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

1] où T et T' sont deux statistiques telles que, dans le premier cas, $P[T<\theta < T'] = 1 - \alpha$ ou, dans le second cas, $P[\theta < T] = 1 - \alpha$ et $P[\theta > T'] = 1 - \alpha$. Nous ne parlerons pas ici des intervalles unilatères, les calculs étant analogues à ceux effectués pour les intervalles bilatères.

3.1.2. Tests d'hypothèses

La construction des tests d'hypothèses répond à la recherche d'une règle de décision. Le problème peut se résumer comme suit : on vise à éprouver une hypothèse, H_0, dont on a de bonnes raisons de penser qu'elle est valide (i.e. elle reste admise jusqu'à preuve du contraire), contre une autre hypothèse, H_1. Ces deux hypothèses sont des affirmations relatives au(x) paramètre(s) (cas paramétrique) ou à la distribution de ceux-ci (cas non paramétrique) et sont mutuellement exclusives. On mène une expérience et en confrontant les observations à la règle de décision, on peut soit rejeter H_0 au profit de H_1, ou au contraire ne pas rejeter (ou encore à conserver), H_0 au profit de H_1. Dans chaque cas, on commet une erreur : si H_0 est vraie et qu'elle est rejetée, on commet alors une certaine erreur α fixée a priori (dite erreur de type I), tandis que si H_1 est vraie et qu'elle est rejetée, on commet alors une certaine erreur β (dite erreur de type II). Les deux hypothèses ne jouent donc pas des rôles symétriques. Les erreurs commises lors de la décision sont reprises ci-dessous.

Décision	H_0 vraie	H_1 vraie
non rejet de H_0	$1-\alpha$	β
rejet de H_1	α	$1-\beta$

L'idéal serait évidemment de trouver un test qui minimiserait à la fois α et β, mais malheureusement, un tel test n'existe pas : pour un échantillon de taille n donnée, si on diminue α, alors automatiquement, β augmente.

3.1.3. Construction de tests

Celle-ci peut s'effectuer de diverses manières : on citera la méthode du rapport de vraisemblance généralisée, la méthode du rapport de vraisemblance monotone, la méthode basée sur la statistique de Rao, la méthode basée sur la statistique de Wald.

Une autre méthode se base sur les intervalles de confiance. Bien que souvent assez facile à mettre en œuvre, il faut insister sur le fait qu'elle n'est qu'une méthode parmi d'autres et par conséquent, la recherche d'intervalles de confiance n'est pas équivalente à celle de la construction de tests. Dans les ouvrages élémentaires, on ne mentionne souvent que cette dernière méthode et cette lacune induit souvent la confusion entre les deux concepts.

3.1.4. Tests basés sur les intervalles de confiance

Dans le problème de test sur une proportion, on considère θ un paramètre réel (inconnu) qui caractérise (partiellement ou entièrement) une distribution statistique et θ_0 un réel connu ; tester l'hypothèse $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ au seuil de signification α peut se faire en déterminant un intervalle de confiance, au niveau de confiance $(1 - \alpha)$, noté $]T_1 ; T_2[$ et à vérifier que $\theta_0 \in]T_1 ; T_2[$. Si c'est le cas, on ne rejette pas H_0, tandis que dans le cas contraire, on rejette H_0.

De même, dans le problème de test sur une différence de deux proportions, si on considère deux populations indépendantes (de tailles respectives n_1 et n_2) telles que chacun des individus de ces deux populations a des probabilités respectivement θ_1 et θ_2 de subir un risque donné, tester $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 > \theta_2$, (ou $H_1' : \theta_1 < \theta_2$ ou encore $H_1'' : \theta_1 \neq \theta_2$) peut se faire en trouvant un intervalle de confiance $]T'_1 ; T'_2[$ au niveau $(1 - \alpha)$ et à vérifier ensuite que $\theta_1 - \theta_2 \in]T'_1 ; T'_2[$. Si c'est le cas, on ne rejettera pas H_0. On rejettera respectivement H_1

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

lorsque $\theta_1 - \theta_2 \leq T'_1$, H_1' lorsque $\theta_1 - \theta_2 \geq T'_2$ et H_1'' lorsque l'une des deux inégalités est satisfaite $\theta_1 - \theta_2 \leq T'_1$ ou $\theta_1 - \theta_2 \geq T'_2$.

3.2. Intervalles de confiance et tests sur le paramètreθ d'une distribution binomiale

3.2.1. Intervalles de confiance bilatères et tests exacts (sans biais)

Soit X une variable aléatoire Bi(n, θ). θ est un réel compris entre 0 et 1 et est inconnu. On souhaite construire un intervalle de confiance bilatère exact (ou sans biais) au niveau de confiance (1 - α) % pour θ. Il s'agit donc de trouver deux statistiques T_1 et T_2 [c'est-à-dire des fonctions de X et des paramètres connus (ici, il n'y a que n)] telles que : $P[T_1 < \theta < T_2] = 1 - \alpha$. L'intervalle de confiance sera donc $]T_1 ; T_2[$. Par intervalle de confiance sans biais, nous entendons que l'égalité précédente est vérifiée.

Ce problème n'est pas évident lorsque la distribution de la statistique suffisante (c'est-à-dire X ici) est discrète, en particulier la distribution binomiale dans le cas que nous considérons.

Il existe plusieurs méthodes conduisant à leur élaboration. Nous en décrivons deux : la méthode « pivotale » et une méthode « alternative ».

Méthode « pivotale »

Cette méthode est celle qui est la plus souvent utilisée et qui est décrite dans les ouvrages « classiques » de statistique élémentaire, mais elle ne conduit donc pas, dans le cas discret, à des intervalles sans biais. Nous en rappelons son principe : il s'agit de trouver une variable aléatoire Y qui ne dépende que des observations et du paramètre (éventuellement vectoriel) à estimer et dont la loi est connue exactement ou asymptotiquement (notée W) et ne dépend pas du paramètre. Il suffit d'écrire alors $P\left[W_{\frac{\alpha}{2}} < Y < W_{1-\frac{\alpha}{2}} \right]$ et ensuite d'isoler le paramètre d'intérêt. Y porte le nom de « variable de pivot » exact ou approximatif, selon le cas. L'existence des quantiles exacts $W_{\frac{\alpha}{2}}$ et $W_{1-\frac{\alpha}{2}}$ requiert évidemment la continuité de la distribution. Elle est donc inapplicable dans le cas qui nous intéresse ici.

Méthode « alternative »

Nous notons X_{obs} la valeur observée de la variable X qui dépend d'un paramètre réel θ inconnu. Soit alors θ une valeur comprise entre 0 et 1. On pose alors $P_\theta [X \leq x]$ la probabilité qu'une variable binomiale Bi (n, θ) soit inférieure ou égale à x.

Proposition

Sous les hypothèses et avec les notations précédentes,

a) il existe deux nombres T_1 et T_2 uniques et compris entre 0 et 1 tels que

$$P_{T_1}[X \geq X_{obs}] = \frac{\alpha}{2} \quad \text{(i)}$$

$$P_{T_2}[X \leq X_{obs}] = \frac{\alpha}{2} \quad \text{(ii)}$$

b) $]T_1, T_2[$ est un intervalle de confiance exact au niveau de confiance α.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

Démonstration :

a) On montre ii) , la démonstration de i) se faisant de manière similaire.

Considérons la fonction suivante : $h : \begin{cases} [0,1] \to [0,1] \\ \theta \quad \to h(\theta) = P_\theta[X \le X_{obs}] \end{cases}$

Cette fonction est continue sur [0, 1], dérivable sur]0, 1[(ceci est trivial) et strictement décroissante (ceci est nettement moins trivial). En outre, $h(0) = 1$ et $h(1) = 0$. Le théorème des valeurs intermédiaires implique alors le résultat. Notons que T_2 ne dépend que de X_{obs} et de n, et est donc bien une statistique.

b) Il suffit de montrer que $P[\theta \ge T_2] = \frac{\alpha}{2}$. On montre de la même façon que $P[\theta \le T_1] = \frac{\alpha}{2}$, ce qui permet de conclure. Mais cela résulte de l'identité entre les deux événements suivants grâce à la définition de h et de T_2 : $\{\theta \ge T_2\} \Leftrightarrow \{P_\theta[X \le X_{obs}] \le \frac{\alpha}{2}\}$

•

Cette proposition permet alors de construire numériquement des intervalles de confiance exacts : T_1 et T_2 s'obtiennent en résolvant numériquement les équations suivantes : $P_{T_2}[X \le X_{obs}] = \frac{\alpha}{2}$ et $P_{T_1}[X \ge X_{obs}] = \frac{\alpha}{2}$, soit dans notre cas précis : $\sum_{k=X_{obs}}^{k=n} C_n^k \theta^k (1-\theta)^{n-k} - \frac{\alpha}{2} = 0$ et $\sum_{k=0}^{k=X_{obs}} C_n^k \theta^k (1-\theta)^{n-k} - \frac{\alpha}{2} = 0$.

La méthode de Newton-Raphson converge rapidement vers les résultats. Par ailleurs, on peut faire usage d'une relation exacte basée sur la distribution de Fisher : au seuil α, la borne inférieure de l'intervalle bilatère est donnée par $\frac{X_{obs}}{X_{obs} + (n - X_{obs} + 1) F_{2(n-X_{obs}+1),\, 2X_{obs},\, 1-\frac{\alpha}{2}}}$, tandis que la borne supérieure est donnée par $\frac{(X_{obs}+1) F_{2(X_{obs}+1),\, 2(n-X_{obs}),\, 1-\frac{\alpha}{2}}}{n - X_{obs} + (X_{obs}+1) F_{2(X_{obs}+1),\, 2(n-X_{obs}),\, 1-\frac{\alpha}{2}}}$.

Le calcul est très facile à réaliser, même avec une machine à calculer scientifique de poche disposant des fonctions statistiques.

Toutefois pour de « grandes » valeurs de n (pratiquement n supérieur à 2000), les méthodes numériques exactes peuvent ne pas fournir de solution satisfaisante. En optimisant les calculs, grâce aux relations de récurrence entre les termes successifs des sommes, on est confronté à des problèmes d'instabilité numérique : ceci se produit par exemple lorsqu'on multiplie des puissances très élevées de quantités proches de 0 par des grands nombres. Il en résulte une instabilité numérique dont les effets sont catastrophiques. Les relations basées sur la distribution de Fisher peuvent également ne pas fournir de résultats valables lorsque le nombre de degrés de liberté est trop grand.

Dans ce cas seulement, il est justifié de se tourner vers des approximations.

3.2.2. Approximations

Les théorèmes et propriétés relatifs aux convergences de variables aléatoires (cf. annexe 2) permettent, sous certaines hypothèses, de faire usage d'approximations. Les propriétés asymptotiques des distributions « lorsque n devient grand » permettent de remplacer, dans certains cas, ces distributions par des distributions normales.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

On trouve dans la littérature un ensemble de conditions plus ou moins restrictives permettant ces approximations. Classiquement, on préconise de faire appel à l'approximation de Poisson lorsque, θ étant suffisamment petit, n est grand et que le produit $n\theta$ reste constant. Les différents auteurs ne sont pas unanimes sur les valeurs de n et de θ. On trouve les inégalités suivantes : $n \geq 20$ et $\theta < 0.05$, $n \geq 100$ et $\theta < 0.05$ ou encore $n \geq 100$ et $n\theta < 10$. En ce qui concerne l'approximation normale, on trouve également diverses conditions sur n et θ : on préconise $n\theta$ et $n(1-\theta)\geq 5$, $n\theta(1-\theta)>9$, $n\theta(1-\theta)\geq 10$, $n > 9\max\left\{\frac{\theta}{(1-\theta)};\frac{(1-\theta)}{\theta}\right\}$ et encore $\theta \pm 2\sqrt{\frac{\theta(1-\theta)}{n}}$ dans [0, 1].

En fait, toutes ces conditions sont valables : tout dépend de la précision sur le résultat que l'on souhaite obtenir (cf. Leemis and Trivedi (1996)), et, au risque de se répéter, il est évident que si on peut éviter de recourir à une approximation en faisant usage des lois exactes (ce qui constitue de moins en moins un obstacle), c'est de loin la solution préférable. Les graphiques repris en annexe permettent de visualiser la qualité des approximations.

Approximation poissonnienne : intervalle de confiance sans biais pour le paramètre d'une loi de Poisson

Si X une variable aléatoire de Poisson de paramètre λ, alors, on montre que $P[X \leq k] = P[Y > 2\lambda]$, où Y est une variable aléatoire $\chi^2_{2(k+1)}$. Dès lors, si on dispose d'un échantillon de taille n, et de moyenne $\overline{x}$, on peut en déduire l'intervalle de confiance bilatère exact pour λ au niveau $(1 - \alpha)$: $\frac{1}{2n}\chi^2_{2n\overline{x};\frac{\alpha}{2}} < \lambda < \frac{1}{2n}\chi^2_{2n(\overline{x}+1);1-\frac{\alpha}{2}}$ où $\chi^2_{k;\alpha}$ est le α-quantile d'une χ^2 à k degrés de liberté. Par conséquent, cette relation facile à mettre en œuvre permet de trouver facilement un intervalle de confiance exact pour le paramètre d'intérêt.

Approximations normales : intervalles de confiance pour la moyenne

Approximation usuelle

Puisqu'on sait que X est asymptotiquement $N\left(\theta,\sqrt{\frac{\theta(1-\theta)}{n}}\right)$, on peut déduire facilement un intervalle de confiance (assez grossier) pour θ. L'intervalle dont les bornes sont données par $\hat{\theta} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$ (où $\hat{\theta}=\frac{x}{n}$ et $z_{\frac{\alpha}{2}}$ est le quantile au niveau $\alpha/2$ de la distribution normale centrée réduite), est donc un intervalle de confiance pour θ au niveau α.

Approximation « améliorée »

En résolvant l'équation du second degré suivante : $(\hat{\theta}-\theta)^2 = z^2_{\frac{\alpha}{2}}\frac{\theta(1-\theta)}{n}$ dans laquelle θ est le paramètre à estimer et $\hat{\theta}$ est la proportion observée, on déduit les bornes suivantes

$$\frac{\left(2\hat{\theta}+\frac{z_{\frac{\alpha}{2}}^2}{n}\right)\pm\sqrt{\frac{z_{\frac{\alpha}{2}}^4}{n^2}+4\hat{\theta}\frac{z_{\frac{\alpha}{2}}^2}{n}\left(1-\hat{\theta}\right)}}{2\left(1+\frac{z_{\frac{\alpha}{2}}^2}{n}\right)}$$

. Elles sont un peu plus précises que les précédentes puisqu'on on ne fait pas appel à une estimation de la variance.

Approximation tenant compte de la correction de continuité

L'approximation normale est meilleure lorsque l'on tient compte d'une correction de continuité qui permet de prendre en compte le fait que l'on remplace une distribution discrète par une distribution continue. Au seuil de signification α, les bornes de l'intervalle de confiance sont données par $\hat{\theta}+\frac{1}{2n}\pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}\left(1-\hat{\theta}\right)}{n}}$. Cet intervalle se déduit de la formule proposée par Yates (cf. infra). Le terme correctif tend évidemment vers 0 lorsque n tend vers l'infini.

Approximation arcsinus

Pour être complet, signalons enfin l'existence d'une approximation basée sur la fonction arcsinus. On peut montrer que, pour de petites valeurs de θ, $\arcsin\left(\sqrt{X}\right)$ est asymptotiquement $N\left(\arcsin\left(\sqrt{\theta}\right), \frac{1}{4n}\right)$. Partant de ce résultat, il est alors facile de construire un intervalle de confiance pour la moyenne θ : il est donné par les bornes suivantes $\sin^2\left(\arcsin\sqrt{\hat{\theta}}\pm z_{\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}\right)$.

Son principal avantage réside dans le fait que ses bornes seront toujours comprises entre 0 et 1 et donc, cette méthode est appropriée lorsqu'on travaille avec des proportions ou des taux.

3.2.3. Comparaison des différents intervalles de confiance

On pourra se référer à l'article de Leemis et Trivedi (1996) mentionné plus haut.

Pour les différentes méthodes, il est possible de calculer l'erreur commise en utilisant les intervalles de confiance obtenus par les différentes méthodes approchées qui ont été présentées ci-dessus. Si on note par $[l_j ; u_j]$ les intervalles trouvés par la méthode « j » où

- j = P dans le cas de l'intervalle de Poisson ;
- j = N dans le cas de l'intervalle de normal simple ;
- j = C dans le cas de l'intervalle de normal tenant compte de la correction de continuité ;
- j = M dans le cas de l'intervalle de normal « amélioré » ;
- j = A dans le cas de l'intervalle trouvé au moyen de l'approximation arcsinus ;
- j = E dans le cas de l'intervalle exact,

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

alors, deux mesures possibles de l'erreur commise, en remplaçant la vraie distribution par une approximation, sont données par $\max\left\{\left|\frac{l_E - l_j}{l_E}\right|; \left|\frac{u_E - u_j}{u_E}\right|\right\}$ et $\max\left\{\left|l_E - l_j\right|; \left|u_E - u_j\right|\right\}$. Il s'agit d'une erreur relative et d'une erreur absolue.

Nous donnons en annexe 2 une représentation graphique de ces deux mesures pour les intervalles de confiance calculés par les différentes méthodes pour différentes valeurs de n et pour différentes valeurs de θ. On trouvera, par ailleurs, différents intervalles de confiance calculés par les différentes méthodes pour n = 10, n = 50, n = 100, pour les valeurs de θ valant 0.1, 0.02 et 0.01.

De manière générale, plus la taille n augmente, plus faibles sont les erreurs et donc meilleures sont les approximations. Dans le cas des approximations normales, ceci est d'autant plus vrai que θ s'approche de la valeur de 0.5. A l'opposé, plus θ s'approche de 0, les approximations de Poisson sont les meilleures.

En ce qui concerne les petites valeurs de n, les choses se présentent différemment : pour les valeurs de θ inférieures à 0.1, l'approximation de Poisson est la meilleure tandis que les approximations basées sur la normale sont dans l'ensemble assez médiocres, sauf peut-être celle qui fait appel à la transformation arc sinus) : elles peuvent conduire à des bornes inférieures qui peuvent être négatives. C'est la cas par exemple, lorsque θ est inférieur à 0.3 et n = 10 ou θ < 0.07 et n = 50. Mais de toutes manières, dans cette situation, l'utilisation de l'intervalle exact est facile et elle évite de la sorte ces problèmes d'approximation.

Lorsque n est « grand », et que θ n'est pas trop petit, par exemple, n ≥ 500 et θ ≥ 0.1, on pourra recourir aux approximations normales. Les erreurs commises sont faibles et on examinera les courbes présentées ci-dessus pour déterminer les erreurs (relative et absolue) que l'on commet.

3.3. Intervalles de confiance et tests sur la différence des paramètres$\delta = \theta 1 - \theta 2$ de deux distributions binomiales

Supposons que l'on souhaite comparer deux proportions et que l'on veuille donc tester

$H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 \neq \theta_2$ ou $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 > \theta_2$

Si X_1 et X_2 sont deux distributions binomiales indépendantes notées respectivement $Bi(n_1, \theta_1)$ et $Bi(n_2, \theta_2)$, alors le problème peut se ramener à celui de la détermination d'intervalles de confiance pour la différence $\delta = \theta_1 - \theta_2$. Les estimateurs de maximum de vraisemblance de θ_1 et θ_2 sont respectivement $\hat{\theta}_1 = \frac{x_1}{n_1}$ et $\hat{\theta}_2 = \frac{x_2}{n_2}$. Dès lors, par le théorème de Zehna, l'estimateur de maximum de vraisemblance de δ est donné par $\hat{\delta} = \frac{x_1}{n_1} - \frac{x_2}{n_2}$. On a que $E\left(\hat{\delta}\right) = \delta$ et $V\left(\hat{\delta}\right) = \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}$. Il suffirait donc de construire les intervalles de confiance exact pour répondre au problème posé, mais ce problème est compliqué. Actuellement, à notre connaissance, il n'existe pas de solution à ce problème.

3.3.1. Intervalles de confiance bilatères basés sur des approximations

La distribution exacte de δ n'est pas connue exactement car c'est une combinaison linéaire de binomiales de paramètres n_1 et n_2 a priori différents. Il n'est donc pas possible de trouver, en général, un intervalle de confiance exact par la méthode pivotale rappelée au

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

paragraphe 3.2.1., ni même par la méthode qualifiée d'alternative. Actuellement, la construction des intervalles de confiance reste basée sur des approximations. Nous en distinguerons trois qui sont toutes basées sur les propriétés asymptotiques des binomiales et sur les propriétés de la loi normale.

Intervalle de confiance normal simple

Selon Dagnelie (STAT1 p. 278), l'approximation de la loi binomiale par la loi normale permet d'écrire, au niveau de confiance (1 - α) %, l'intervalle de confiance sur δ donné par les bornes $\hat{\delta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_1\left(1-\hat{\theta}_1\right)}{n_1}+\frac{\hat{\theta}_2\left(1-\hat{\theta}_2\right)}{n_2}}$.

Intervalle de confiance corrigé (Yates)

Il est possible d'améliorer la précision de l'intervalle précédent en tenant compte d'une correction de continuité qui prend en compte le fait que l'on fait usage d'une distribution discrète à la place d'une distribution continue.

Yates propose l'intervalle suivant : $\hat{\delta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_1\left(1-\hat{\theta}_1\right)}{n_1}+\frac{\hat{p}_2\left(\hat{\theta}-\hat{\theta}_2\right)}{n_2}+\left\{\frac{1}{2n_1}+\frac{1}{2n_2}\right\}}$

Intervalle de confiance « amélioré » (Hauck et Anderson)

En utilisant l'estimateur non biaisé de la variance de la différence δ (et ceci se justifie d'autant plus que les populations sont de petite taille), Hauck et Anderson proposent, quant à eux, l'intervalle amélioré suivant : $\hat{\delta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_1\left(1-\hat{\theta}_1\right)}{n_1}+\frac{\hat{p}_2\left(\hat{\theta}-\hat{\theta}_2\right)}{n_2}+\left\{\frac{1}{2\min\left(n_1,n_2\right)}\right\}}$

Les relations proposées précédemment permettent de construire immédiatement un intervalle de confiance approximatif et donc de tester la différence des proportions.

Le problème de la recherche d'un intervalle de confiance exact pour la différence entre deux paramètres de binomiales, problème simple en première apparence, est donc loin d'être évident et actuellement, ce problème n'a pas encore trouvé, à notre connaissance, de réponse. Il y a encore une recherche à poursuivre dans ce domaine. En attendant cette réponse, nous préconisons de recourir à une des deux approximations prenant en compte une correction de continuité (Yates ou Hauck et Anderson).

3.3.2. Test exact

Si on souhaite seulement tester l'hypothèse $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 \neq \theta_2$, alors, pour de « petites valeurs de n_1 et n_2, il existe une solution.

Dans un article récent, Berger (1996) propose un test qui répond à la question.

Soit le test $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 < \theta_2$.

Si X et Y sont deux binomiales Bi(m,θ_1) et Bi(n,θ_2),

Si $Z(x,y) = \frac{\hat{\theta}_2 - \hat{\theta}_1}{\sqrt{\hat{\theta}\left(1-\hat{\theta}\right)\left(\frac{1}{m}+\frac{1}{n}\right)}}$ où $\hat{\theta} = \frac{x+y}{m+n}$,

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

Alors la p - valeur du test est égale à $p_Z = \sup \sum_{(a,b)\in R_Z(x,y)} C_m^a p^a (1-p)^{m-a} C_n^b p^b (1-p)^{n-b}$,

où le sup est pris sur l'ensemble des p compris entre 0 et 1,

et $R_Z(x,y) = \{(a,b) \in \{0,1,\ldots m\} \times \{0,1,\ldots n\}$ tels que $Z(a,b) > Z(x,y)\}$.

On rejette H_0 ssi p_Z est inférieure au seuil α fixé au départ.

Partant de ce résultat, il suffit de tester :

1) $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 < \theta_2$. On déterminera une p-valeur p_{Z1}

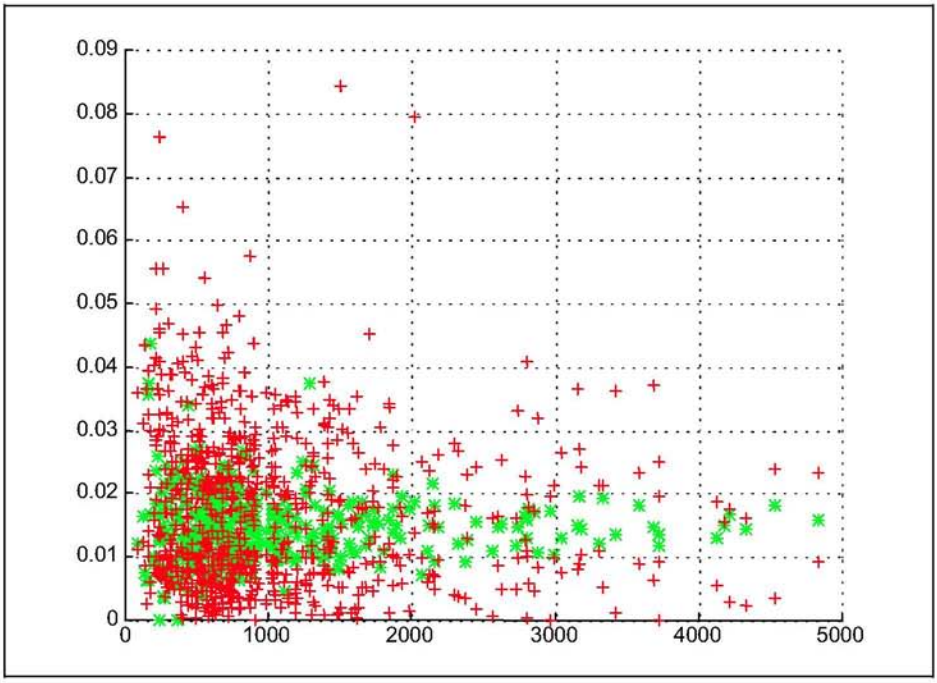
2) $H_0 : \theta_2 = \theta_1$ contre $H_1 : \theta_2 < \theta_1$. On déterminera une p-valeur p_{Z2}

On rejettera l'hypothèse ssi $\min(p_{Z1}, p_{Z2})$ est inférieure au seuil α.

4. Application à des données réelles

4.1. Intervalles de confiance sur des taux bruts de mortalité

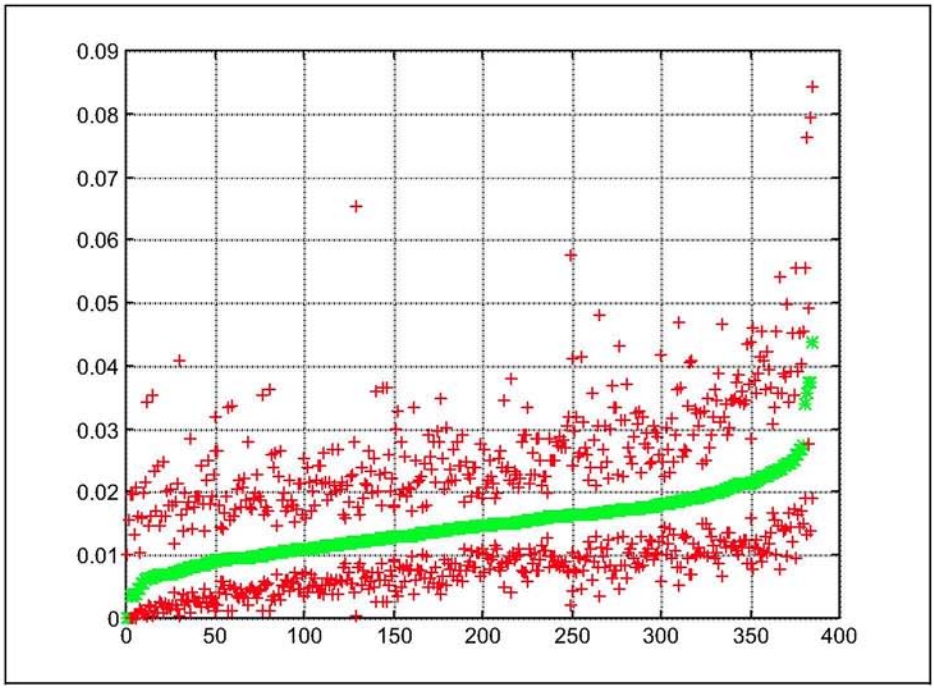
Pour chacun des taux bruts de mortalité observés en 1910 sur 385 communes, la figure suivante donne en fonction de la taille de la population le taux brut de mortalité observé (indiqués par *) et les bornes des intervalles de confiance exactes (indiquées par +) au niveau de confiance de 95 %.



On notera la tendance à la décroissance des intervalles de confiance en fonction de la taille de la population soumise au risque : en moyenne, plus cette taille est petite, plus l'intervalle correspondant à une longueur importante et inversement. Par ailleurs, ces taux bruts de mortalité fluctuent autour d'une valeur proche de 0.015.

En ordonnant les taux bruts de mortalité par valeurs croissantes, les mêmes données se présentent comme suit :

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.



Il est intéressant de constater que pour des taux bruts de mortalité proches, les intervalles de confiance peuvent être très différents. Ceci s'explique par l'effet de la taille de la population.

4.2. Intervalles de confiance sur le nombre d'événements

On se propose dans l'exemple suivant de calculer les taux bruts de mortalité de quelques quartiers de la ville de Namur entre 1991 et 1996 et d'en déduire un intervalle de confiance pour le nombre de décès. Nous retenons les quartiers de Wierde, de Jambes centre et de Loyers, ainsi que l'ensemble de la ville de Namur.

Pour ces quartiers le tableau ci-dessous donne la population moyenne, le nombre de décès qui ont été observés sur la période, le taux brut de mortalité observé, les bornes de l'intervalle de confiance exact pour le taux brut de mortalité et le nombre de décès correspondant à ces bornes.

Quartier	Pop. moyenne	Nbre décès	TBM	Borne inf.	Borne sup.	Décès inf.	Décès sup.
Wierde	443	24	0.05418	0.03502	0.07954	15 (15)	35 (33)
Jambes centre	5811	428	0.07365	0.06707	0.08067	390 (399)	469 (467)
Loyers	1452	38	0.02617	0.01859	0.03575	27 (26)	52 (50)
Namur[9] (total)	104502	6430	0.06153	0.06004	0.06305	6274 (6277)	6589 (6582)

Les valeurs numériques obtenues avec les méthodes faisant appel aux approximations conduisent à des résultats différents. On a repris entre parenthèses les nombres de décès

[9] Pour l'ensemble de la ville de Namur, l'intervalle exact ne peut être calculé, car on se heurte à des problèmes numériques. Vu la faible valeur du taux brut de mortalité, et la taille élevée, on se tournera vers l'approximation de Poisson qui présente l'avantage de permettre un calcul d'intervalle exact.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

obtenus en faisant usage de l'approximation normale simple. On peut apprécier le gain de précision qui est obtenu en recourant à l'intervalle exact.

Dans le cas de Wierde, la prise en compte de la correction de continuité mène à un nombre de décès compris entre 15 et 34, tandis que l'approximation basée sur la transformation arcsinus donne 16 et 34 décès et l'approximation de Poisson fournit 15 et 36 décès. Ces différences absolues sont faibles (une unité en plus ou en moins), mais l'erreur relative commise est relativement importante (de l'ordre de 3 à 4 %).

4.3. Intervalles de confiance sur la différence de taux bruts de mortalité

La question qui était posée au départ était la suivante : « est-ce que les niveaux de mortalité sont (significativement) différents entre deux quartiers ? ». Les deux méthodes exposées ci-dessus (test exact et intervalle de confiance) permettent maintenant répondre à cette question.

Pour ce faire, nous déterminerons un intervalle de confiance à 95 % sur la différence entre les deux taux bruts de mortalité observés. La taille des populations étant assez grande, le calcul se fait en utilisant la méthode de Hauck et Anderson. Si l'intervalle trouvé contient la valeur nulle, alors, la conclusion est qu'il n'y a pas de différence significative, donc égalité, entre les deux taux.

En reprenant les mêmes quartiers de la ville de Namur que précédemment, on détermine les intervalles de confiance suivants :

Quartier	Wierde	Jambes centre	Loyers	Namur (total)
Wierde	- -			
Jambes centre	**[-0.0405 0.0038]**	- -		
Loyers	[0.0065 0.0518]	[0.0372 0.0584]	- -	
Namur (total)	**[-0.0274 0.0149]**	[0.0053 0.0191]	[-0.0434 -0.0267]	- -

Les cellules en gras indiquent des couples de quartiers qui présentent des taux bruts de mortalité significativement non différents, puisque les intervalles de confiance contiennent la valeur 0, tandis que les autres présentent des différences significatives.

La comparaison des taux bruts de mortalité de Wierde ($n = 443$, $d = 24$, $\theta = 0.05418$) et de Loyers ($n = 1452$, $d = 38$ et $\theta = 0.02617$) à l'aide du test proposé par Berger conduit aux résultats numériques suivants :

$H_0 : TBM_{Wierde} = TBM_{Loyers}$
$H_1 : TBM_{Wierde} < TBM_{Loyers}$

$p_{z1} = 0.9976$

$H_0 : TBM_{Loyers} = TBM_{Wierde}$
$H_1 : TBM_{Loyers} < TBM_{Wierde}$

$p_{z2} = 0.0044$

Les conclusions du test sont donc : pour le premier test, on rejette H_1 avec quasi certitude (et ceci était prévisible) tandis que pour le second test, on rejette H_0 pour tout seuil inférieur à 0.0044. En d'autres termes, on accepte avec un risque très minime de se tromper l'hypothèse de différence non nulle entre les deux taux. (On retrouve ainsi le résultat obtenu approximativement au moyen du test basé sur l'intervalle de confiance, ce qui est rassurant !).

Il y a donc bien une différence significative entre les deux taux bruts de mortalité. Les deux quartiers ont donc des mortalités significativement différentes, et dans une étape

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

ultérieure, il pourrait être utile d'en rechercher la cause, ce qui nécessiterait évidemment des informations supplémentaires.

Remarque

Il importe de manipuler les égalités trouvées avec beaucoup de précautions et de prendre attention de tirer des conclusions abusives : ces égalités sont des relations statistiques et donc toujours entachées d'une erreur. Ce ne sont pas des égalités stricto sensu et ne vérifient pas toutes les propriétés usuelles. En particulier, la transitivité n'est pas satisfaite (i.e. si a = b et si b = c alors, on ne peut affirmer que a = c). Ainsi, Wierde et Jambes centre ont le même taux brut de mortalité et il en est de même pour Wierde et Namur (total), mais Jambes centre et Namur (total) ont des taux bruts de mortalité significativement différents.

5. Conclusions

A partir des mesures de comptage des occurrences d'un phénomène démographique régi par une loi binomiale de paramètre θ inconnu, nous avons déterminé un intervalle de confiance exact sur le paramètre de cette loi : chaque mesure de comptage peut donc être affectée d'une précision, et ce pour n'importe quelle taille de population.

En outre, nous pouvons comparer le niveau d'un phénomène démographique sur deux espaces différents grâce à des méthodes exactes ou approchées.

Les méthodes de calcul présentées sont facilement et directement applicables à partir de données réelles. Seule la contrainte numérique peut encore éventuellement constituer un obstacle, auquel cas, il est opportun de se tourner vers les approximations : pour des événements rares (θ petit), il est recommandé de travailler avec l'approximation poissonnienne (pour laquelle il est facile de trouver des intervalles exacts), tandis que pour des événements non rares, on optera pour les intervalles déterminés à l'aide de l'approximation normale tenant compte d'une correction de continuité.

Pour la comparaison de deux paramètres de binomiales, les méthodes approchées permettent de procéder aux tests usuels. La méthode proposée par Hauck et Anderson présente l'avantage de donner des résultats acceptables pour des valeurs faibles de θ_1 et θ_2. Selon la précision requise, on fera appel à l'une ou à l'autre méthode. Un test exact a été proposé par Berger, mais il se révèle assez lourd à mettre en oeuvre pour les situations mettant en jeu des populations de taille « importante ».

Les méthodes développées ci-dessus sont prometteuses et ouvrent une voie de généralisation possible à d'autres paramètres démographiques usuels tels l'espérance de vie, l'indice conjoncturel de fécondité, …. Ces méthodes doivent encore être développées, mais nous pensons au stade actuel qu'elles constitueront une étape qui donnera à la démographie la possibilité de se libérer de la contrainte des « grands nombres » et d'analyser statistiquement l'étude des populations de tailles restreintes.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

BIBLIOGRAPHIE

BERGER, R.L. (1996), More Powerful Tests From Confidence Interval p Values, *The American Statistician*, 50, 4, 314 - 318.

BICKEL, P.J. and DOKSUM, K.A. (1977), *Mathematical Statistics*, Holden - Day, Inc, Oakland, California, 492 p.

DAGNELIE, P. (1992), *Statistique Théorique et Appliquée,* tome 1, Presses Agronomiques de Gembloux, 492 p.

HAUCK, W. W. and ANDERSON, S. (1986), A Comparison of Large-Sample Confidence Interval Methods For The Difference of Two Binomial Probabilities Distributions, *The American Statistician*, 40, 1, 318 - 322.

LEEMIS, L. M. and TRIVEDI, K. S. (1996), A Comparison of Approximate Intervals Estimators for the Bernoulli Parameter, *The American Statistician*, 50, 1, 63 - 68.

MOOD, A.M., GRAYBILL, F.A. and BOES, D.C. (1974), *Introduction to the Theory of Statistics*, Mac Graw - Hill, 564 p.

SAPORTA, G. (1990), *Probabilités Analyse des données et Statistique*, Technip, PARIS, 493 p.

TASSI, P. (1989), *Méthodes statistiques*, Economica, PARIS, 474 p.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

ANNEXE 1 : Distributions statistiques usuelles

1.1. Distribution binaire (ou indicatrice, ou de Bernoulli)

C'est la loi d'une variable aléatoire X qui peut prendre deux valeurs 1 (succès) ou 0 (échec) avec les probabilités respectives θ et $1 - \theta$ ($0 \leq \theta \leq 1$). On montre immédiatement que l'espérance, $E(X) = \theta$ et la variance, $V(X) = \theta(1 - \theta)$. Cette distribution est appropriée pour décrire et étudier les proportions et les taux. Par exemple, dans une population infinie, dont certains individus possèdent une caractéristique donnée, un individu tiré au hasard possède la caractéristique ($X = 1$) ou il ne la possède pas ($X = 0$).

1.2. Distribution binomiale

On répète dans des conditions identiques (succès ou échec et probabilité de succès constante au fil des épreuves) n expériences indépendantes de Bernoulli. En d'autres termes, on observe n fois l'occurrence (succès) ou la non occurrence (échec) d'un événement pour lequel la probabilité d'obtention d'un succès est constante à chaque expérience et vaut θ. Ce processus est connu sous le nom de « schéma de Bernoulli ».

Si on note par X la variable aléatoire « nombre de succès au cours des n expériences », alors X est la somme de n variables de Bernoulli X_i indépendantes et de même paramètre θ. Par définition, X est une variable binomiale notée Bi(n, p) et on sait alors que la probabilité d'observer k succès au cours des n expériences est donnée par $P[X = k] = C_n^k \theta^k (1-\theta)^{n-k}$ (avec $k = 0, 1, \ldots n$ et $C_n^k = \frac{n!}{k!(n-k)!}$ où $k! = k(k - 1)\ldots 2$ et par convention $0! = 1$). Par ailleurs, on montre que $E(X) = n\theta$ et $V(X) = n\theta(1 - \theta)$. Par exemple, le tirage au hasard, dans une population infinie, de n individus possédant chacun l'un des deux caractères opposés et le prélèvement de n individus dans une population finie, avec remise au fur et à mesure des prélèvements, correspondent à ce schéma.

La distribution binomiale suppose donc deux hypothèses :

- l'indépendance des expériences donnant chacune lieu à deux résultats exclusifs (succès ou échec) ;
- la constance de la probabilité de « succès » au cours des n expériences.

1.3. Distribution hypergéométrique

Soit une urne contenant N billes dont R sont rouges et B sont bleues, et supposons que l'on procède à un tirage sans remise[10] de n billes parmi ces R+B billes. On considère alors la variable aléatoire X qui prend pour valeurs le nombre de billes rouges extraites parmi les n. Le tirage s'effectuant sans remise, il n'y a pas indépendance entre les extractions et on a alors que $P[X = k] = \frac{C_R^k C_B^{n-k}}{C_N^n}$. On a alors $E(X) = n\frac{R}{N}$ et $V(X) = n\frac{R}{N}\frac{B}{N}\frac{N-n}{N-1}$. Lorsque N est grand que le rapport $\frac{R}{N}$ tend vers une constante θ, alors on montre facilement que la loi hypergéométrique tend vers la loi binomiale de paramètres N etθ.

1.4. Distribution de Poisson

Si X désigne le nombre de survenances d'un phénomène aléatoire dans un hypervolume V d'un espace vectoriel tel que :

- la probabilité que plus d'une survenance se produise dans un hypervolume dont la mesure tend vers 0, est négligeable devant la probabilité qu'une seule survenance se produise dans cet hypervolume

[10] Si par contre le tirage se fait avec remise, on se retrouve dans les conditions d'application des hypothèses qui définissent la loi binomiale de paramètres R+B, θ où θ est connu et vaut $\frac{R}{N}$.

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

- la probabilité qu'une seule survenance se produise dans un hypervolume V donné est au premier ordre proportionnel à la mesure de cet hypervolume,
- il y a indépendance entre les survenances,

Alors, $P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$ où λ est une constante positive qui dépend de la mesure de V. X suit alors une loi de Poisson et on note dans ce cas $X \sim P_0(\lambda)$. X modélise l'occurrence des événements « rares ». On a immédiatement $E(X) = \lambda$ et $V(X) = \lambda$.

Un des principaux avantages de cette distribution réside dans le fait que sa moyenne coïncide exactement avec sa variance et que celles-ci sont indépendantes de la taille de la population étudiée. De plus, la somme de deux variables indépendantes de Poisson de paramètres respectifs λ_1 et λ_2 est une variable de Poisson de paramètre $\lambda_1+\lambda_2$. ce qui n'est pas le cas pour la somme de deux binomiales de paramètres θ_1 et θ_2 différents.

1.5. Distribution normale (ou de Gauss)

Lorsqu'on effectue une mesure d'une quantité qui peut prendre une valeur réelle quelconque, le résultat de cette mesure est toujours entaché d'une erreur. Par exemple, la mesure de la position d'un point, ou de la longueur, ou de la masse d'un corps physique, … est toujours connue avec une précision finie. Au niveau macroscopique, tous les instruments de mesure ont une limite à la précision qu'ils peuvent fournir, tandis qu'au niveau microscopique, ceci découle immédiatement du principe d'incertitude d'Heisenberg. Si on répète un très grand nombre d'opérations de mesure, on observe que ces mesures se distribuent autour d'une valeur centrale et présentent une dispersion plus ou moins importante autour de cette valeur centrale. Ce constat permet de définir une variable gaussienne. On dira qu'une variable aléatoire X suit une distribution normale de paramètres μ ($\in$ R) et σ ($\in R^+$) lorsque $\forall$ a $\in$ R : $P[X \leq a] = \int_{-\infty}^{a} f(t)dt$ où $f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$. On note alors $X \sim N(\mu; \sigma^2)$. On a alors immédiatement $E(X) = \mu$ et $Var(X) = \sigma^2$. Cette distribution joue un rôle capital en probabilité et en statistique.

On consultera, par exemple Dagnelie, Saporta ainsi que Mood et al. pour une description des propriétés de ces distributions.

ANNEXE 2 : Convergences des variables aléatoires

Les convergences des variables aléatoires présentent un grand intérêt à la fois théorique et pratique, car lorsque la taille de l'échantillon est « grande », elles permettent, sous certaines conditions assez larges, de faire abstraction de la distribution de ces variables aléatoires et de considérer que celles-ci sont régies par une loi normale. Sans entrer dans les détails, nous rappelons ci-dessous les principaux théorèmes, ainsi que les principales propriétés qu'on en déduit.

2.1. Loi faible des grands nombres

Soient X_i ($i = 1, \ldots, n$) n variables aléatoires indépendantes de moyennes respectives μ_i, et de variances respectives σ_i^2, supposées finies.

Si $\frac{1}{n}\sum_{i=1}^{i=n}\mu_i \xrightarrow[n\to+\infty]{} \mu$ et $\frac{1}{n^2}\sum_{i=1}^{i=n}\sigma_i^2 \xrightarrow[n\to+\infty]{} 0$, alors $\frac{1}{n}\sum_{i=1}^{i=n}X_i = \bar{X}_n \xrightarrow{P} \mu$.

(C'est la convergence dite « faible » ou en probabilité, c'est-à-dire $\forall\varepsilon>0, \forall\eta>0 \exists n_0 \in N$ tel que $\forall n>n_0 : P\left[\left|\bar{X}_n - \mu\right| > \varepsilon\right] < \eta$)

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

2.2. Loi forte des grands nombres

Soient X_i (i = 1, …, n) n variables aléatoires indépendantes de moyennes respectives μ_i, et de variances respectives σ_i^2.

Si les μ_i et les σ_i^2 sont telles que $\frac{1}{n}\sum_{i=1}^{i=n}\mu_i \xrightarrow[n\to\infty]{} \mu$ et $\sum_{i=1}^{i=n}\frac{\sigma_i^2}{i^2}$ est une série convergente, alors $\frac{1}{n}\sum_{i=1}^{i=n}X_i = \bar{X}_n \xrightarrow{p.s.} \mu$. (C'est la convergence forte ou presque sûre, ce qui signifie que $P\left[\left\{\omega \text{ tels que } \lim_{n\to\infty}\bar{X}_n(\omega) \neq \mu(\omega)\right\}\right] = 0$: l'ensemble des points pour lesquels l'égalité n'est pas satisfaite est de mesure nulle)

2.3. Théorème central - limite

Soient X_i (i = 1, …, n) n variables aléatoires indépendantes de même loi (éventuellement inconnue) de moyenne μ et de variance σ^2, alors $Y_n = \frac{1}{\sqrt{n}}\frac{\sum_{i=1}^{i=n}X_i - n\mu}{\sigma} \xrightarrow{L} N(0;1)$ (C'est la convergence en loi : elle exprime ici la convergence, en tout point x réel, des fonctions de répartition F_{Y_n} vers la fonction de répartition d'une variable aléatoire normale réduite).

Propriétés

Il découle immédiatement de ces théorèmes précédents les propriétés suivantes :

- Si Xn est une suite de variables aléatoires binomiales Bi(n, □n) telles que le produit nθn $\to \lambda$ avec $\lambda > 0$ lorsque $n \to +\infty$, alors $X_n \xrightarrow{L} P_0(\lambda)$.
- Si Xn est une suite de variables aléatoires binomiales Bi(n, θ) alors $\frac{X_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{L} N(0;1)$.
- Si Xn est une suite de variables aléatoires de Poisson P0(□n) telles que λn$\to\lambda$, alors $\frac{X_n - \lambda}{\sqrt{\lambda}} \xrightarrow{L} N(0;1)$.

L'intérêt pratique de ces propriétés est évident : dès que la taille d'un échantillon est « suffisamment » grande, et pour peu que les conditions d'application soient satisfaites, on peut approximer la loi binomiale, soit par la loi de Poisson, soit par la loi normale. De plus, si n est assez grand, on peut sous certaines conditions approximer la loi de Poisson par une loi normale.

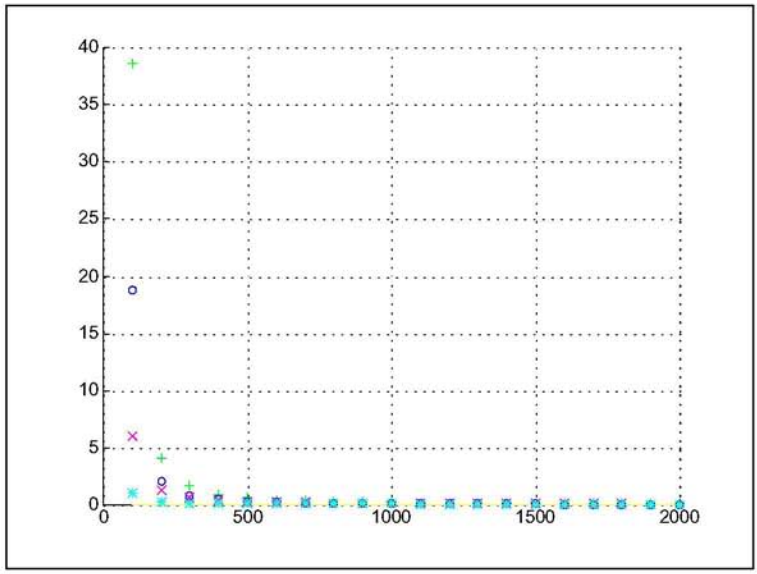
ANNEXE 3 : Qualité des approximations - intervalles de confiance exact et approches

Variation de $\max\left\{\left|l_E - l_j / l_E\right|; \left|u_E - u_j / u_E\right|\right\}$ pour θ = 0,01 et n variant de 100 à 2000 par pas de 100.

α = 5 %

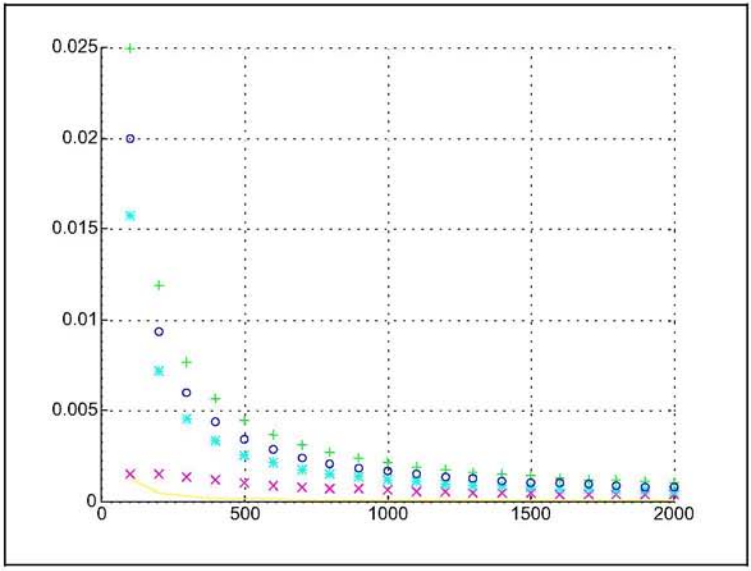
(En trait continu : intervalle de Poisson, + : intervalle normal simple o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et arcsinus).

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.



Variation de $\max\left\{\left|l_E - l_j\right|; \left|u_E - u_j\right|\right\}$ pour $\theta = 0,01$ et n variant de 100 à 2000 par pas de 100. $\alpha = 5\%$.

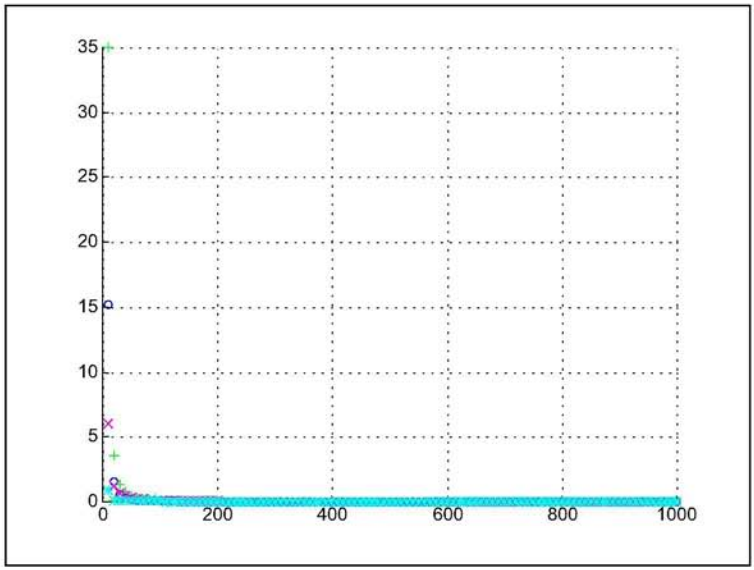
(En trait continu : intervalle de Poisson, + : intervalle normal simple o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et arcsinus).



Variation de $\max\left\{\left|\frac{l_E - l_j}{l_E}\right|; \left|\frac{u_E - u_j}{u_E}\right|\right\}$ pour $\theta = 0,1$ et n variant de 10 à 1000 par pas de 10. $\alpha = 5\%$.

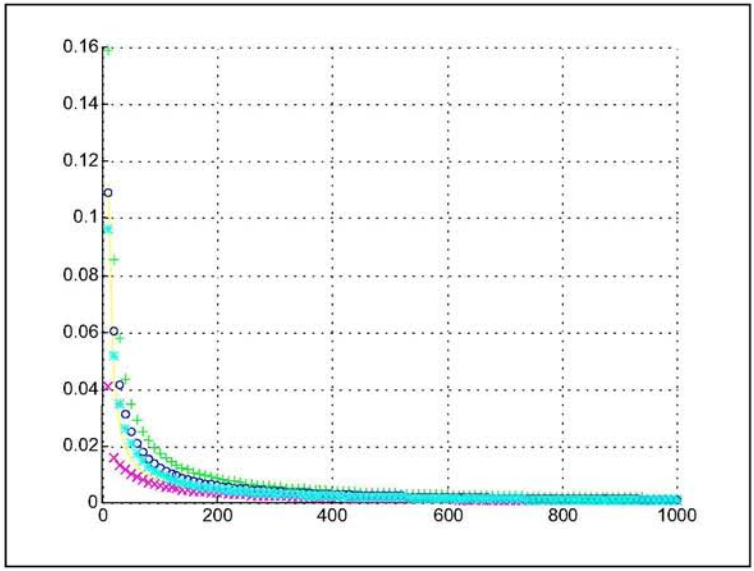
(En trait continu : intervalle de Poisson, + : intervalle normal simple o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et arcsinus).

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.



Variation de $\max\left\{\left|l_E - l_j\right|;\ \left|u_E - u_j\right|\right\}$ pour $\theta = 0{,}1$ et n variant de 10 à 1000 par pas de 10. $\alpha = 5\ \%$.

(En trait continu : intervalle de Poisson, + : intervalle normal simple o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et approximation arcsinus).

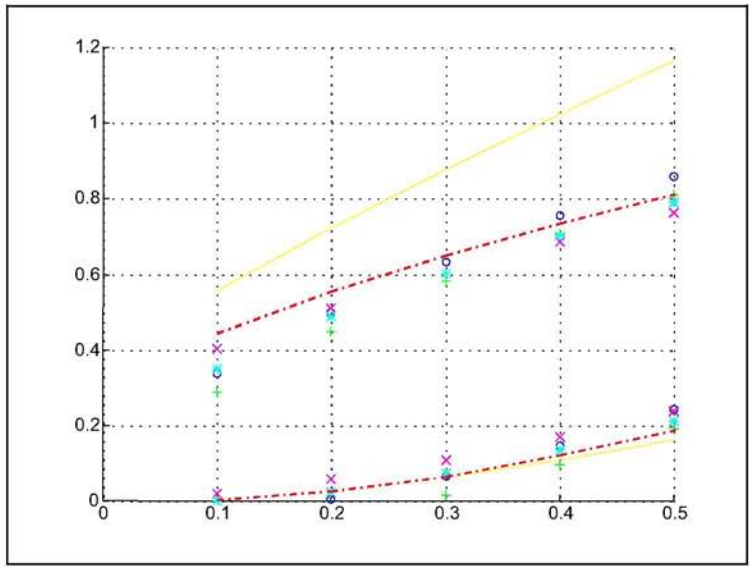


Intervalles de confiance (α = 5 %) pour le paramètre d'une binomiale : méthode exacte et approximations pour n = 10 et θ varie de 0.1 à 0.5 par pas de 0.1

En trait pointillés : intervalle exact, en trait continu : intervalle de Poisson, + : intervalle normal simple

o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et arcsinus.

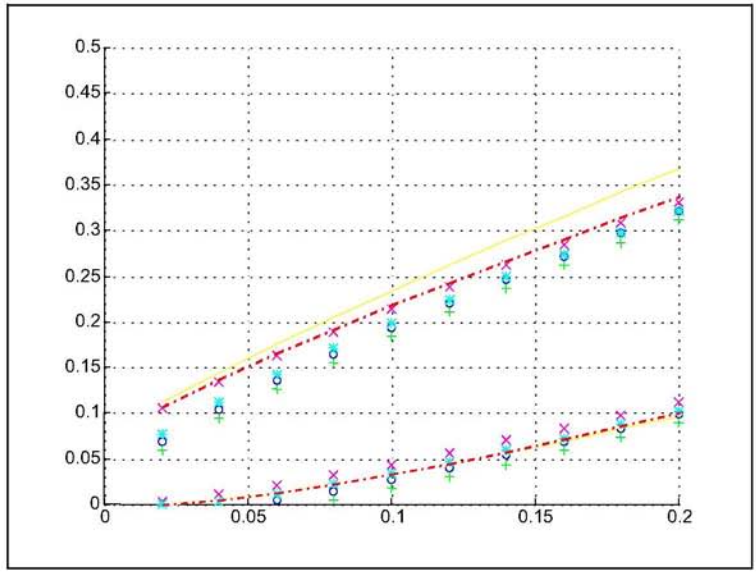
AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.



Intervalles de confiance (α = 5 %) pour le paramètre d'une binomiale : méthode exacte et approximations pour n = 50 et θ varie de 0.02 à 0.2 par pas de 0.02.

En trait pointillés : intervalle exact, en trait continu : intervalle de Poisson, + : intervalle normal simple

o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et arcsinus.

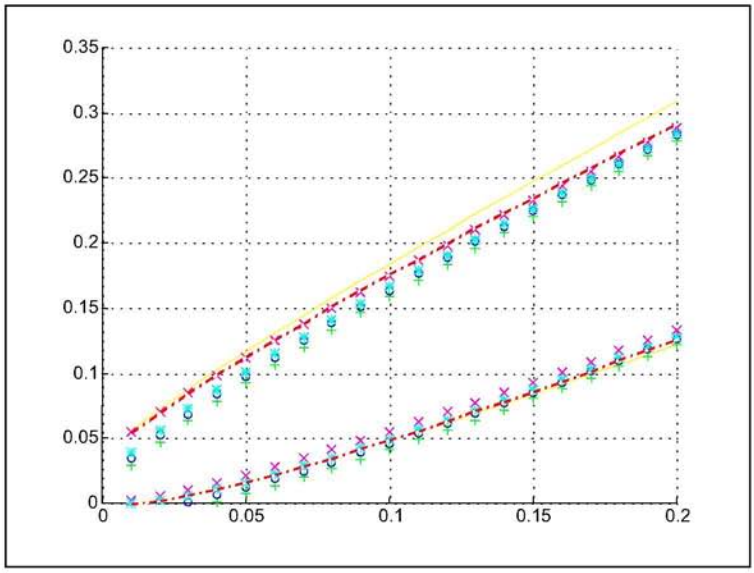


Intervalles de confiance (α = 5 %) pour le paramètre d'une binomiale : méthode exacte et approximations pour n = 100 et θ varie de 0.01 à 0.2 par pas de 0.01

En trait pointillés : intervalle exact, en trait continu : intervalle de Poisson, + : intervalle normal simple

AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.

o : intervalle normal avec correction de continuité, x : intervalle normal « amélioré » ; * : intervalle normal et arcsinus.



AIDELF. De Bartolo Guiseppe, Poulain Michel. 2003. La démographie appliquée à la gestion publique et des entreprises. Association internationale des démographes de langue française, ISBN : 2-950-9356-4-8, 218 pages.