Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

IS Math314 **Année 2009**

Corrigé de l'examen du 12 juin 2009

Ex 1. *Intervalles de confiance pour une probabilité inconnue (4 points)*

On dispose d'un 400-échantillon observé $(X_i(\omega))_{1\le i\le 400}$ d'une loi de Bernoulli. Le paramètre p de cette loi est inconnu, mais on sait avec certitude que $0 < p < 0{,}3$. Le nombre de valeurs 1 dans l'échantillon observé est 72.

1) La variable aléatoire X_1 suivant la loi de Bernoulli de paramètre p, a pour variance $p(1-p)$. Or la fonction $f : p \mapsto p(1-p)$ est croissante sur $[0, 1/2]$, donc le meilleur majorant (c'est-à-dire le plus petit) de $p(1-p)$ sur $]0, 0{,}3]$ est $f(0{,}3)$. Ainsi

$$\operatorname{Var} X_1 = p(1-p) \le 0{,}3(1-0{,}3) = 0{,}21. \tag{1}$$

2) Posons $S_n = X_1 + \cdots + X_n$ et $\overline{X} = n^{-1}S_n$. Le théorème de de Moivre-Laplace (ou le théorème limite central) appliqué aux variables i.i.d. de Bernoulli X_i de carré intégrable et d'espérance $\mathbf{E}X_i = p$ nous donne :

$$S_n^* = \sqrt{\frac{n}{p(1-p)}}(\overline{X} - p) \xrightarrow[n\to+\infty]{\text{loi}} Z,$$

avec Z de loi gaussienne standard $\mathfrak{N}(0, 1)$. Cette convergence en loi justifie l'approximation suivante pour $n = 400$ considéré comme « grand ».

$$\forall t > 0, \quad P\left(|S_n^*| \le t\right) \simeq P(|Z| \le t) = 2\Phi(t) - 1.$$

Ceci s'écrit aussi

$$P\left(\overline{X} - \frac{t\sqrt{p(1-p)}}{\sqrt{400}} \le p \le \overline{X} + \frac{t\sqrt{p(1-p)}}{\sqrt{400}}\right) \simeq 2\Phi(t) - 1. \tag{2}$$

En choisissant t tel que $2\Phi(t) - 1 = 0{,}95$, soit $t = 1{,}96$, en négligeant l'erreur d'approximation gaussienne et en utilisant la majoration (1), on en déduit :

$$P\left(\overline{X} - \frac{1{,}96\sqrt{0{,}21}}{20} \le p \le \overline{X} + \frac{1{,}96\sqrt{0{,}21}}{20}\right) \ge 0{,}95.$$

Dans l'échantillon observé, il y a 72 valeurs 1, les autres étant toutes nulles, donc $\overline{X} = \frac{72}{400} = 0{,}18$. Ceci nous amène à proposer comme intervalle de confiance I au niveau 95% pour p :

$$I = [\,0{,}135; 0{,}225\,].$$

3) On note $\overline{X}$ et S^2 la moyenne et la variance empiriques de l'échantillon $(X_i)_{1\leq i\leq 400}$. En utilisant la formule de Koenig pour la variance empirique, on a

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2.$$

Comme X_i est à valeurs dans $\{0,1\}$, on a $X_i^2 = X_i$ puisque $0^2 = 0$ et $1^2 = 1$. L'égalité ci-dessus se réécrit alors

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i - \left(\overline{X}\right)^2 = \overline{X} - \left(\overline{X}\right)^2 = \overline{X}(1-\overline{X}).$$

Ainsi l'estimateur de $\operatorname{Var} X_1 = p(1-p)$ obtenu en remplaçant p par son estimateur « naturel » $\overline{X}$ n'est autre que la variance empirique. Un conséquence sympathique est que l'on économise ici le calcul de la somme des X_i^2 pour obtenir S^2.

4) Le théorème limite central avec autonormalisation appliqué aux variables aléatoires i.i.d. X_i de carré intégrable et de variance non nulle nous donne :

$$\sqrt{\frac{n}{S^2}}\left(\overline{X}-p\right) = \sqrt{\frac{n}{\overline{X}(1-\overline{X})}}\left(\overline{X}-p\right) \xrightarrow[n\to+\infty]{\text{loi}} Z,$$

avec Z de loi gaussienne standard $\mathfrak{N}(0,1)$. Ceci légitime le remplacement dans (2) de la variance inconnue $p(1-p)$ par $\overline{X}(1-\overline{X})$, ce qui nous donne

$$P\left(\overline{X} - \frac{t\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{400}} \leq p \leq \overline{X} + \frac{t\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{400}}\right) \simeq 2\Phi(t) - 1.$$

Passant à l'application numérique avec $\overline{X} = 0{,}18$ et $t = 1{,}96$, nous obtenons l'intervalle de confiance J au niveau 95% pour p :

$$I = [\,0{,}142; 0{,}218\,].$$

Ex 2. *Matrices de covariance et couples gaussiens (6 points)*

En dimension 1, il est facile de vérifier que tout nombre réel positif t est la variance d'une certaine variable aléatoire. Par exemple prenons X variable de Bernoulli de paramètre $1/2$, sa variance est $1/4$. Pour toute constante c, $\operatorname{Var}(cX) = c^2 \operatorname{Var} X = c^2/4$. En prenant $c = 2\sqrt{t}$ et $Y = cX$, on obtient $\operatorname{Var} Y = t$. On a choisi ici pour X une v.a. de Bernoulli par souci de simplicité, mais n'importe quelle variable aléatoire de variance non nulle aurait fait l'affaire. La question analogue en dimension 2 est beaucoup moins triviale et s'énonce : « étant donné une matrice 2×2, à quelles conditions est-elle matrice de covariance d'un certain vecteur aléatoire ? ».

1) Pour chercher des conditions *nécessaires*, supposons d'abord que M est effectivement la matrice de covariance d'un certain vecteur aléatoire (X,Y) de $\mathbb{R}^2$. Alors

$$M = \begin{pmatrix} \operatorname{Var} X & \operatorname{Cov}(X,Y) \\ \operatorname{Cov}(Y,X) & \operatorname{Var} Y \end{pmatrix}.$$

Comme la variance est positive, la diagonale principale de M est formée par les réels positifs $a = \operatorname{Var} X$ et $b = \operatorname{Var} Y$. La covariance étant symétrique, on a $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X) = c$. D'autre part avec ces notations, l'inégalité de Cauchy-Schwarz pour les covariances :

$$|\operatorname{Cov}(X,Y)|^2 \leq \operatorname{Var} X \operatorname{Var} Y$$

s'écrit $c^2 \leq ab$. Finalement M est de la forme :

$$M = \begin{pmatrix} a & c \\ c & b \end{pmatrix}, \qquad a, b \in \mathbb{R}^+,\ c \in \mathbb{R}, \quad ab - c^2 \geq 0. \tag{3}$$

Réciproquement, soit M une matrice 2×2 de la forme (3), nous allons montrer qu'elle est la matrice de covariance d'un certain couple aléatoire (X,Y) de $\mathbb{R}^2$.

2) Examinons d'abord le cas particulier où $ab = 0$. Puisque $0 \leq c^2 \leq ab$, on en déduit immédiatement que $c = 0$. Il y a trois possibilités.

1. $a = 0$ et $b = 0$, alors M est la matrice nulle, c'est la matrice de covariance d'un vecteur aléatoire constant (C_1, C_2) puisque la variance d'une constante est nulle.
2. $a = 0$ et $b > 0$, alors $M = \begin{pmatrix} 0 & 0 \\ 0 & b \end{pmatrix}$ est la matrice de covariance d'un vecteur aléatoire (C_1, Y), où C_1 désigne n'importe quelle constante réelle et Y n'importe quelle variable aléatoire de variance b (d'après l'introduction il en existe toujours une).
3. $a > 0$ et $b = 0$, alors $M = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$ est la matrice de covariance d'un vecteur aléatoire (X, C_2), où C_2 désigne n'importe quelle constante réelle et X n'importe quelle variable aléatoire de variance a.

Dans chacun de ces trois sous-cas, on a utilisé implicitement le fait que si l'une des composantes d'un couple aléatoire est constante, la covariance lorsqu'elle existe[1] est nulle. C'est une conséquence immédiate de l'inégalité de Cauchy-Schwarz ci-dessus et de la nullité de la variance d'une constante.

3) On suppose désormais que ab est non nul. On pose alors

$$\sigma_1 = \sqrt{a},\ \sigma_2 = \sqrt{b}, \quad \rho = \frac{c}{\sqrt{ab}} \quad (\text{donc } \rho \in [-1,1]),$$

on se donne un couple (U,V) de variables aléatoires réelles *indépendantes*, de carré intégrable et de variance 1 et on définit l'application linéaire g par :

$$g : \mathbb{R}^2 \to \mathbb{R}^2, \quad (u,v) \longmapsto (x,y), \text{ avec } \begin{cases} x = \sigma_1 u, \\ y = \sigma_2 \rho u + \sigma_2 \sqrt{1-\rho^2}\, v. \end{cases} \tag{4}$$

Les composantes X et Y du vecteur aléatoire $(X,Y) = g(U,V)$ sont de carré intégrable comme combinaisons linéaires des variables aléatoires U et V de carré intégrable. Par

1. c'est-à-dire si l'autre composante est de carré intégrable.

conséquent (X, Y) possède une matrice de covariance. Pour la calculer, on se ramène à U et V en rappelant que la variance d'une somme de deux v.a. indépendantes est la somme des variances et que la covariance de deux v.a. indépendantes est nulle.

$$\operatorname{Var} X = \operatorname{Var}(\sigma_1 U) = \sigma_1^2 \operatorname{Var} U = \sigma_1^2 = a. \tag{5}$$

$$\begin{aligned}\operatorname{Var} Y &= \operatorname{Var}\left(\sigma_2\rho U + \sigma_2\sqrt{1-\rho^2}\, V\right)\\ &= \sigma_2^2\rho^2 \operatorname{Var} U + 2\sigma_2^2\rho\sqrt{1-\rho^2}\operatorname{Cov}(U,V) + \sigma_2^2(1-\rho^2)\operatorname{Var} V\\ &= b\rho^2 + 0 + b(1-\rho^2) = b.\end{aligned} \tag{6}$$

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \operatorname{Cov}\left(\sigma_1 U, \sigma_2\rho U + \sigma_2\sqrt{1-\rho^2}\, V\right)\\ &= \sigma_1\sigma_2\rho \operatorname{Var} U + \sigma_1\sigma_2\sqrt{1-\rho^2}\operatorname{Cov}(U,V)\\ &= \sqrt{a}\sqrt{b}\frac{c}{\sqrt{ab}} + 0 = c.\end{aligned} \tag{7}$$

Au vu de (5), (6) et (7), la matrice de covariance K du vecteur aléatoire (X, Y) est :

$$K = \begin{pmatrix} \operatorname{Var} X & \operatorname{Cov}(X,Y) \\ \operatorname{Cov}(Y,X) & \operatorname{Var} Y \end{pmatrix} = \begin{pmatrix} a & c \\ c & b \end{pmatrix} = M.$$

Nous avons ainsi construit un vecteur aléatoire (X, Y) ayant M pour matrice de covariance.

Tout ce qui précède montre qu'une matrice 2×2 à coefficients réels est la matrice de covariance d'un vecteur aléatoire de $\mathbb{R}^2$ *si et seulement si* elle est de la forme (3).

4) Lorsque U et V ci-dessus sont gaussiennes $\mathfrak{N}(0,1)$, le vecteur aléatoire (U, V) est *gaussien* en raison de *l'indépendance* de U et V. Cette indépendance nous permet aussi d'affirmer que le couple (U, V) admet une densité $f_{(U,V)}$ donnée par

$$f_{(U,V)}(u,v) = f_U(u) f_V(v) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-v^2}{2}\right) = \frac{1}{2\pi}\exp\left(-\frac{u^2+v^2}{2}\right).$$

Comme (X, Y) est image du vecteur gaussien (U, V) par l'application *linéaire* g, c'est encore un vecteur gaussien.

En prime nous avons obtenu une méthode simple de simulation d'un vecteur aléatoire gaussien de $\mathbb{R}^2$ de matrice de covariance donnée.

5) Soit (X', Y'), un vecteur aléatoire gaussien d'espérance $(0, 0)$ et de matrice de covariance K. Puisque K est une matrice de covariance, elle est de la forme (3), avec $a = \sigma_1^2$, $b = \sigma_2^2$, $c = \operatorname{Cov}(X', Y')$. Le déterminant de K n'étant pas nul, le cas $ab = 0$ est exclu (car $\det K = ab - c^2 \geq 0$ donc si $ab = 0$, c est nul et $\det K = 0$). Les variances de X' et Y' sont donc strictement positives et le coefficient de corrélation ρ est bien défini :

$$\rho = \frac{\operatorname{Cov}(X', Y')}{\sqrt{\operatorname{Var} X' \operatorname{Var} Y'}} = \frac{c}{\sqrt{ab}}.$$

Le vecteur aléatoire gaussien $(X, Y) = g(U, V)$ construit aux deux questions précédentes à partir de U et V i.i.d. $\mathfrak{N}(0, 1)$ et de la matrice de covariance $M = K$ a pour espérance

$(0,0)$ et pour matrice de covariance K. Il a donc *même loi* que (X',Y'). On sait que si g est linéaire bijective, $g(U,V)$ a une densité qui se déduit de celle de (U,V) par la formule de changement de variable linéaire pour les densités. Pour vérifier la bijectivité de g, on note que

$$\det g = \sigma_1\sigma_2\sqrt{1-\rho^2} = \sqrt{ab(1-\rho^2)}$$

et

$$1-\rho^2 = \frac{ab-c^2}{ab}$$

d'où

$$\det g = \sqrt{\det K} \neq 0.$$

D'après la fomule de changement de variable linéaire pour les densités, (X,Y) admet pour densité :

$$f_{X,Y}(x,y) = \frac{1}{|\det g|} f_{U,V}(g^{-1}(x,y)).$$

L'application linéaire inverse de g se calcule en résolvant le système (4) en considérant u et v comme des inconnues :

$$\begin{aligned} u &= \frac{x}{\sigma_1} \\ v &= \frac{-\sigma_2\rho x + \sigma_1 y}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \end{aligned}$$

Pour alléger les écritures, calculons séparément :

$$\begin{aligned} u^2+v^2 &= \frac{x^2}{\sigma_1^2} + \frac{(-\sigma_2\rho x+\sigma_1 y)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)} \\ &= \frac{\sigma_2^2(1-\rho^2)x^2 + \sigma_2^2\rho^2x^2 + \sigma_1^2y^2 - 2\rho\sigma_1\sigma_2xy}{\sigma_1^2\sigma_2^2(1-\rho^2)} \\ &= \frac{\sigma_2^2x^2 + \sigma_1^2y^2 - 2\rho\sigma_1\sigma_2xy}{\sigma_1^2\sigma_2^2(1-\rho^2)} \end{aligned}$$

Finalement, la densité de (X,Y) est donnée par

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{\sigma_2^2x^2+\sigma_1^2y^2-2\rho\sigma_1\sigma_2xy}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right).$$

Comme (X',Y') a même loi que (X,Y), il admet lui aussi pour densité $f_{X,Y}$.

6) Vérifions finalement que si $\det K = 0$, soit le vecteur aléatoire (X',Y') est p.s. constant égal à $(0,0)$, soit sa loi est portée par une droite D, c'est-à-dire $P((X',Y') \in D) = 1$. En notant encore $a=\sigma_1^2$, $b=\sigma_2^2$, $c=\mathrm{Cov}(X',Y')$, l'hypothèse s'écrit $ab=c^2$.

Nous avons déjà examiné le cas $ab=0$ à la question 2), revoyons le en prenant en compte le caractère gaussien de (X',Y').

1. Si $a = b = 0$, alors le vecteur gaussien constant $(0,0)$ a pour espérance $(0,0)$ et pour matrice de covariance K. Comme le loi gaussienne de (X', Y') est entièrement déterminée par son vecteur espérance et sa matrice de covariance, (X', Y') a *même loi* que le vecteur constant $(0,0)$ et en particulier toute la masse de sa loi est portée par le point $(0,0)$, ce qui s'écrit aussi $P((X', Y') = (0,0)) = 1$.
2. Si $a = 0$ et $b > 0$, considérons le vecteur $(0, Y)$ où Y est gaussienne $\mathfrak{N}(0, \sigma_2)$. Ses composantes sont gaussiennes et indépendantes puisqu'une v.a. constante est toujours indépendante de n'importe quelle v.a. y compris d'elle même[2]. Donc $(0, Y)$ est gaussien. Comme il a même vecteur espérance et même matrice de covariance que (X', Y'), ces deux vecteurs gaussiens ont même loi. En particulier $P(X' = 0) = P((X', Y') \in \{0\} \times \mathbb{R}) = P((0, Y) \in \{0\} \times \mathbb{R}) = 1$, donc (X', Y') appartient avec probabilité 1 à la droite $D = \{0\} \times \mathbb{R}$ d'équation $x = 0$.
3. Si $a > 0$ et $b = 0$, on vérifie de même que (X', Y') appartient avec probabilité 1 à la droite d'équation $y = 0$.

Il nous reste à étudier le cas $ab = c^2 > 0$. On peut alors définir le coefficient de corrélation $\rho = c(ab)^{-1/2}$ et noter que nécessairement $\rho = \pm 1$. L'application linéaire g définie par (4) se réécrit plus simplement :

$$g : \mathbb{R}^2 \to \mathbb{R}^2, \quad (u, v) \longmapsto (x, y), \text{ avec } \begin{cases} x = \sigma_1 u, \\ y = \sigma_2 \rho u. \end{cases}$$

Elle n'est plus bijective ($\det g = 0$) et elle envoie $\mathbb{R}^2$ sur la droite D d'équation

$$y = \frac{\sigma_2 \rho}{\sigma_1} x.$$

En notant encore (U, V) un vecteur gaussien à composantes indépendantes et de loi $\mathfrak{N}(0,1)$, nous savons que le vecteur aléatoire $(X, Y) = g(U, V)$ est gaussien d'espérance $\mathbf{E}g(U, V) = g(\mathbf{E}U, \mathbf{E}V) = g(0,0) = (0,0)$ et de matrice de covariance K, donc de même loi que (X', Y'). En particulier,

$$P((X', Y') \in D) = P((X, Y) \in D) = P(g(U, V) \in D) = 1.$$

Ex 3. *Estimation de la moyenne d'une loi de Poisson mélange (10 points)*

Pour les assurances, le nombre N de sinistres causés en une année par un assuré donné est souvent modélisé par une loi de Poisson de paramètre α. On a alors $\mathbf{E}N = \alpha$, ce qui permet d'interpréter le paramètre α comme un nombre moyen de sinistres. Ce modèle ne colle pas bien à la réalité lorsque l'on étudie l'ensemble des assurés correspondant à une police donnée (par exemple les conducteurs de véhicule d'un certain type dans une certaine zone géographique). Pour prendre en compte l'hétérogénéité des assurés, on remplace alors α par $R\alpha$ où R est une variable aléatoire positive d'espérance 1, appelée niveau de risque relatif. Le modèle « bons risques – mauvais risques » en assurance automobile est le plus simple de ce type. Il revient à partager l'ensemble d'assurés étudié

2. Si vous l'ignoriez, vérifiez le en exercice.

en « bons conducteurs » pour lesquels le nombre de sinistres suit la loi de Poisson de paramètre $r_1\alpha$ (avec $r_1 < 1$) et en « mauvais conducteurs » pour lesquels le nombre de sinistres suit la loi de Poisson de paramètre $r_2\alpha$ avec $r_2 > 1$. Le problème est qu'il est difficile de décider au vu de l'historique des accidents qui est bon ou mauvais conducteur. Parmi les assurés n'ayant pas causé de sinistre lors des années précédentes, il peut très bien y avoir des mauvais conducteurs « chanceux » et parmi les assurés ayant causé au moins un sinistre, il peut tout aussi bien y avoir des bons conducteurs « malchanceux ». On considère donc que le niveau de risque relatif d'un assuré choisi au hasard est une variable aléatoire R vérifiant :

$$P(R = r_1) = p_1 > 0,\ P(R = r_2) = p_2 > 0, \quad p_1 + p_2 = 1, \quad \mathbf{E}R = 1. \tag{8}$$

Les paramètres p_i, r_i peuvent être estimés par l'observation approfondie des comportements au volant d'un échantillon d'assurés. Dans toute la suite, nous les supposerons connus. Le seul paramètre inconnu de notre modèle sera donc α. On s'intéresse au nombre N de sinistres causés par un assuré choisi au hasard (on ne sait pas s'il est bon ou mauvais conducteur).

1) La variable aléatoire n étant à valeurs dans $\mathbb{N}$, sa loi est déterminée par les $P(N = k)$, $k \in \mathbb{N}$. Pour calculer la probabilité de l'évènement $N = k$, il suffit d'appliquer la formule des probabilités totales en conditionnant par les deux cas possibles « bon risque » $(R = r_1)$ et « mauvais risque » $(R = r_2)$, ces deux évènements complémentaires étant de probabilité non nulle puisque $p_1 > 0$ et $p_2 > 0$:

$$\begin{aligned} P(N = k) &= P(N = k \mid R = r_1)P(R = r_1) + P(N = k \mid R = r_2)P(R = r_2) \\ &= p_1 \frac{\mathrm{e}^{-r_1\alpha}(r_1\alpha)^k}{k!} + p_2 \frac{\mathrm{e}^{-r_2\alpha}(r_2\alpha)^k}{k!}. \end{aligned} \tag{9}$$

On dit que N suit la loi de *Poisson mélange* de moyenne α et de niveau de risque relatif R donné par (8).

2) Pour calculer l'espérance et la variance (sous réserve d'existence) de la variable aléatoire N, on rappelle d'abord que si X suit une loi de Poisson de paramètre β,

$$\mathbf{E}X = \sum_{k=0}^{+\infty} kP(X = k) = \sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-\beta}k\beta^k}{k!} = \beta \tag{10}$$

et

$$\operatorname{Var} X = \sum_{k=0}^{+\infty} k^2 P(X = k) - \beta^2 = \sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-\beta}k^2\beta^k}{k!} - \beta^2 = \beta, \tag{11}$$

cette dernière formule pouvant aussi s'écrire

$$\mathbf{E}X^2 = \beta + \beta^2. \tag{12}$$

Calculons maintenant l'espérance de la v.a. positive discrète N.

$$\mathbf{E}N = \sum_{k=0}^{+\infty} kP(N = k) = \sum_{k=0}^{+\infty} k \left(p_1 \frac{\mathrm{e}^{-r_1\alpha}(r_1\alpha)^k}{k!} + p_2 \frac{\mathrm{e}^{-r_2\alpha}(r_2\alpha)^k}{k!} \right).$$

Les deux termes de la somme entre parenthèses étant positifs (pour chaque k)[3], ceci peut encore s'écrire :

$$\begin{aligned}\mathbf{E}N &= p_1 \sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-r_1\alpha} k (r_1\alpha)^k}{k!} + p_2 \sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-r_2\alpha} k (r_2\alpha)^k}{k!} \\ &= p_1 r_1 \alpha + p_2 r_2 \alpha \quad \text{(en appliquant (10) avec } \beta = r_1\alpha,\ \beta = r_2\alpha) \\ &= \alpha(p_1 r_1 + p_2 r_2) \\ &= \alpha \mathbf{E}R = \alpha.\end{aligned}$$

On note au passage que l'espérance de la v.a. positive N est finie, ce qui justifie son intégrabilité. Pour vérifier l'existence de $\operatorname{Var} N$, on commence par calculer $\mathbf{E}N^2$ pour voir si cette quantité est finie.

$$\begin{aligned}\mathbf{E}N^2 &= \sum_{k=0}^{+\infty} k^2 P(N=k) \\ &= \sum_{k=0}^{+\infty} k^2 \left(p_1 \frac{\mathrm{e}^{-r_1\alpha}(r_1\alpha)^k}{k!} + p_2 \frac{\mathrm{e}^{-r_2\alpha}(r_2\alpha)^k}{k!} \right) \\ &= p_1 \sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-r_1\alpha} k^2 (r_1\alpha)^k}{k!} + p_2 \sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-r_2\alpha} k^2 (r_2\alpha)^k}{k!} \quad \text{(somme de séries à termes } \textit{positifs}) \\ &= p_1(r_1\alpha + r_1^2\alpha^2) + p_2(r_2\alpha + r_2^2\alpha^2) \quad \text{(en appliquant (11) et (12) avec } \beta = r_i\alpha) \\ &= \alpha(p_1 r_1 + p_2 r_2) + \alpha^2(p_1 r_1^2 + p_2 r_2^2) \\ &= \alpha + \alpha^2 \mathbf{E}R^2.\end{aligned}$$

On vérifie ainsi que $\mathbf{E}N^2$ est fini et on en déduit :

$$\operatorname{Var} N = \mathbf{E}N^2 - (\mathbf{E}N)^2 = \alpha + \alpha^2 \mathbf{E}R^2 - \alpha^2 = \alpha + \alpha^2(\mathbf{E}R^2 - 1) = \alpha + \alpha^2\big(\mathbf{E}R^2 - (\mathbf{E}R)^2\big).$$

Finalement :

$$\mathbf{E}N = \alpha, \tag{13}$$

$$\operatorname{Var} N = \alpha + \alpha^2 \operatorname{Var} R. \tag{14}$$

On remarque que la loi de N est plus dispersée autour de α qu'une loi de Poisson classique de paramètre α dont l'espérance et la variance valent α.

3) On considère désormais le modèle statistique $(\Omega, \mathcal{F}, P_\alpha)_{\alpha\in]0,+\infty[}$ et un n-échantillon $N_1, \ldots, N_n$ associé. Autrement dit, pour toute valeur de α, les variables aléatoires N_i sont P_α-indépendantes et de même loi sous P_α donnée par (9), en remplaçant cela va de soi, P par P_α dans cette formule. On note $\overline{N}$ la moyenne empirique de cet échantillon. Compte-tenu du calcul d'espérance et de variance que nous venons d'achever, la loi forte

3. Pourquoi cette remarque ?

des grands nombres et le théorème limite central nous fournissent respectivement les convergences suivantes :

$$\forall \alpha > 0, \quad \overline{N} \xrightarrow[n\to+\infty]{P_\alpha-\text{p.s.}} \alpha; \tag{15}$$

$$\forall \alpha > 0, \quad \sqrt{\frac{n}{\operatorname{Var}_\alpha N}}\big(\overline{N} - \alpha\big) \xrightarrow[n\to+\infty]{P_\alpha-\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{16}$$

4) On se propose maintenant de justifier la convergence en loi suivante :

$$\forall \alpha > 0, \quad T_n := \sqrt{\frac{n}{\frac{1}{n} + \overline{N}(1+\rho^2\overline{N})}}\big(\overline{N} - \alpha\big) \xrightarrow[n\to+\infty]{P_\alpha-\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{17}$$

où l'on a posé $\rho^2 = \operatorname{Var} R$ (la loi de R est connue et ne dépend pas de α).

La convergence presque-sûre étant conservée par addition [4] et par image continue, il découle de (15) que :

$$\forall \alpha > 0, \quad \sqrt{\frac{1}{\frac{1}{n} + \overline{N}(1+\rho^2\overline{N})}} \xrightarrow[n\to+\infty]{P_\alpha-\text{p.s.}} \sqrt{\frac{1}{\operatorname{Var}_\alpha N}}. \tag{18}$$

Le terme $\frac{1}{n}$ au dénominateur n'est là que pour empêcher que ce dénominateur s'annule et éviter ainsi une discussion pénible sur ce cas (de probabilité non nulle). Ceci nous incite à réécrire la v.a. T_n dans (17) sous la forme

$$T_n = \underbrace{\sqrt{\frac{\operatorname{Var}_\alpha N}{\frac{1}{n} + \overline{N}(1+\rho^2\overline{N})}}}_{=:V_n} \times \underbrace{\sqrt{\frac{n}{\operatorname{Var}_\alpha N}}\big(\overline{N} - \alpha\big)}_{=:W_n}.$$

Par (18), V_n converge P_α-p.s. vers 1, tandis que par (16), W_n converge en loi sous P_α vers Z gaussienne $\mathfrak{N}(0,1)$. Le lemme de Slutsky nous assure alors de la convergence en loi sous P_α du *couple aléatoire* (V_n, W_n) vers $(1, Z)$. Cette convergence en loi entraîne celle du produit V_nW_n vers Z en prenant l'image du couple (V_n, W_n) par l'application produit $\mathbb{R}^2 \to \mathbb{R}$, $(x,y) \mapsto xy$ qui est *continue*. Toutes ces convergences ayant lieu pour tout $\alpha > 0$, ceci termine la vérification de (17).

5) La convergence en loi (17) légitime pour n grand l'approximation gaussienne :

$$P_\alpha(|T_n| \leq 1{,}96) \simeq P(|Z| \leq 1{,}96) = 2\Phi(1{,}96) - 1 = 0{,}95.$$

4. La conservation de la convergence p.s. par addition est ici un argument de paresseux pour voir que si X_n converge p.s. vers X, alors $1/n + X_n$ converge p.s. vers X puisque $1/n$ converge p.s. (trivialement) vers 0. On peut aussi le vérifier en revenant à la définition de la convergence presque sûre. En détail, pour justifier (18), partant de la convergence p.s. de $\overline{N}$ vers α, on utilise successivement la continuité au point α de $t \mapsto t(1+\rho^2 t)$, l'addition avec $1/n$ et finalement la continuité de $s \mapsto s^{-1/2}$ au point $\operatorname{Var}_\alpha N = \alpha(1+\rho^2\alpha) > 0$.

En négligeant l'erreur due à cette approximation gaussienne et en résolvant en encadrement de α l'inégalité $|T_n| \leq 1{,}96$, on en déduit l'intervalle de confiance I au niveau 95% :

$$I = \left[\overline{N} - 1{,}96\sqrt{\frac{\frac{1}{n} + \overline{N}(1+\rho^2\overline{N})}{n}} \; ; \; \overline{N} + 1{,}96\sqrt{\frac{\frac{1}{n} + \overline{N}(1+\rho^2\overline{N})}{n}} \right].$$

Notons que les bornes de cet intervalle sont explicitement calculables à partir des observations *via* le calcul de $\overline{N}$. Si on avait utilisé la convergence (16) pour proposer un intervalle à bornes aléatoires contenant α avec une probabilité de 95% (toujours en négligeant l'erreur d'approximation gaussienne), on aurait obtenu l'intervalle :

$$J = \left[\overline{N} - 1{,}96\sqrt{\frac{\mathrm{Var}_\alpha\, N}{n}} \; ; \; \overline{N} + 1{,}96\sqrt{\frac{\mathrm{Var}_\alpha\, N}{n}}. \right].$$

L'ennui c'est que les bornes de J ne sont pas calculables à partir des seules observations : elles dépendent de la quantité inconnue $\mathrm{Var}_\alpha\, N$. Ce n'est donc pas un intervalle de confiance.

6) Le tableau ci-dessous donne un échantillon observé de taille 100.

0	1	0	2	1	0	0	2	0	0	0	0	1	0	0	1	0	0	2	0
0	1	1	0	0	0	2	0	0	1	0	1	0	1	1	0	0	0	2	1
1	0	0	2	0	0	1	1	1	0	1	0	0	1	0	0	0	1	0	0
0	3	1	0	0	0	0	1	1	1	1	0	0	1	0	0	1	0	0	0
1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	2	0	3

a) Il y a 4 valeurs distinctes dans ce tableau. Pour obtenir la fréquence de chacune d'elles, on compte son nombre d'apparitions dans l'échantillon et on le divise par la taille de l'échantillon. Ceci nous donne les résultats quivants :

Valeur	0	1	2	3
Fréquence	0,61	0,30	0,07	0,02

Ce tableau de fréquences nous permet de dessiner la fonction de répartition empirique associée à cet échantillon, cf. figure 1, p. 11.

b) Le tableau des fréquences permet de simplifier le calcul de la moyenne empirique $\overline{N}(\omega) = \frac{1}{100}\sum_{i=1}^{100} N_i(\omega)$ en notant que dans les valeurs observées $N_1(\omega), \ldots, N_{100}(\omega)$, il n'y a que 4 valeurs distinctes $j = 0, 1, 2, 3$ de fréquence respective f_j de sorte qu'en regroupant toutes les valeurs égales on a :

$$\overline{N}(\omega) = \sum_{j=0}^{3} j f_j = 0{,}61 \times 0 + 0{,}30 \times 1 + 0{,}07 \times 2 + 0{,}02 \times 3 = 0{,}5$$

c) On sait que $p_1 = 0{,}25$ (donc $p_2 = 0{,}75$), $r_1 = 0{,}7$ et $r_2 = 1{,}1$. En se rappelant que $\mathbf{E}R = 1$, on peut alors calculer

$$\mathrm{Var}\, R = \mathbf{E}R^2 - (\mathbf{E}R)^2 = p_1 r_1^2 + p_2 r_2^2 - 1 = 0{,}03.$$

d) On peut maintenant expliciter l'intervalle de confiance I calculé à partir de cet échantillon :

$$I = [0{,}359; 0{,}641]\,.$$

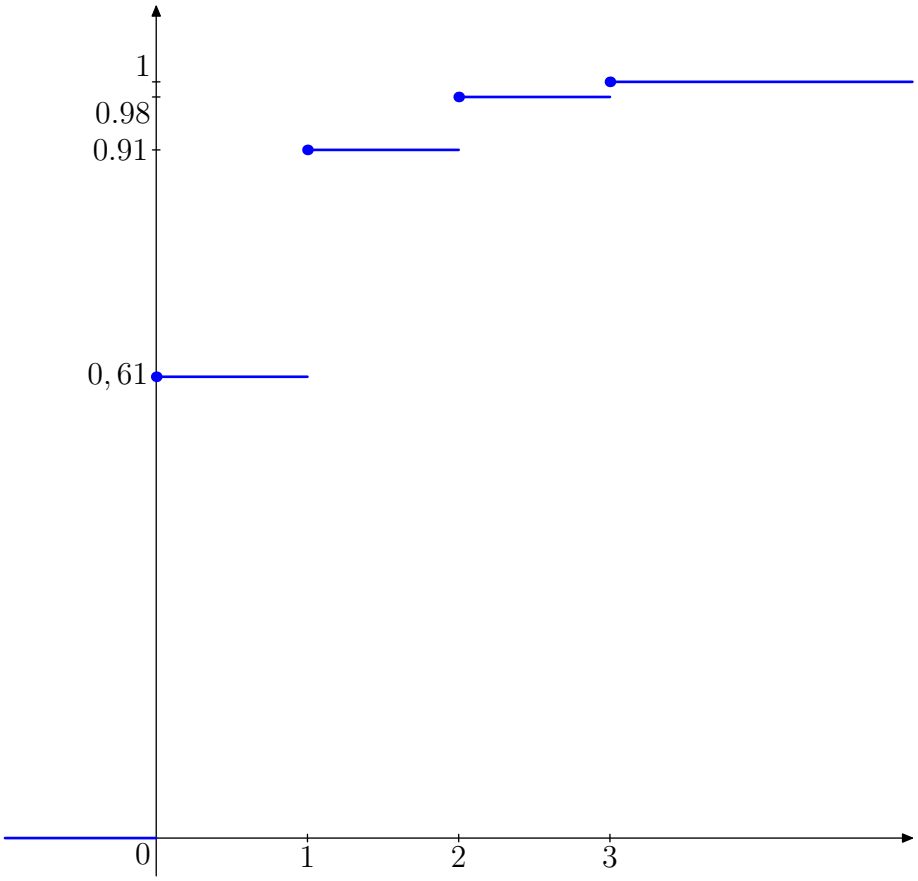

FIG. 1 – F.d.r. empirique associée à l'échantillon