HEC II 2005

L'objet du problème est d'étudier quelques propriétés d'un estimateur du paramètre p d'une loi géométrique.

Partie I. Formule du binôme négatif.

Pour tout couple (n,r) d'entiers naturels tels que $0 \leqslant r \leqslant n$, on rappelle la formule du " triangle de Pascal" : $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$

1. Montrer que pour tout entier r de $[\![1, n]\!]$, on a : $\binom{n}{r} = \sum\limits_{k=r}^{n} \binom{k-1}{r-1}$.

2. Soit (n,r) un couple d'entiers naturels, tels que $1 \leqslant r \leqslant n$. Pour tout réel x de $]0,1[$, on définit la fonction $f_{r,n}$ par : $f_{r,n}(x) = \sum\limits_{k=r}^{n} \binom{k}{r} x^k$.

   a) Montrer, pour tout réel x de $]0;1[$, l'égalité : $(1-x)f_{r,n}(x) = xf_{r-1,n-1}(x) - \binom{n}{r}x^{n+1}$.

   b) On suppose l'entier r fixé. Montrer, lorsque n tend vers $+\infty$, l'équivalence : $\binom{n}{r} \underset{n\to+\infty}{\sim} \dfrac{n^r}{r!}$.

3. Soit x un réel fixé de $]0;1[$ et soit r un entier naturel fixé. On veut établir l'existence de la limite de $f_{r,n}(x)$ lorsque n tend vers $+\infty$, et déterminer la valeur de cette limite.

   a) Justifier l'existence et donner la valeur de $\lim\limits_{n\to+\infty} f_{0,n}(x)$ et $\lim\limits_{n\to+\infty} f_{1,n}(x)$.

   b) Soit r un entier naturel non nul. On suppose que, pour tout réel x de $]0,1[$, on a :

   $$\lim_{n\to+\infty} f_{r-1,n}(x) = \frac{x^{r-1}}{(1-x)^r}.$$

   Montrer que, pour tout réel x de $]0,1[$, $\lim\limits_{n\to+\infty} f_{r,n}(x) = \dfrac{x^r}{(1-x)^{r+1}}$. Ainsi, $\sum\limits_{k=r}^{+\infty} \binom{k}{r}x^k = \dfrac{x^r}{(1-x)^{r+1}}$.

Partie II. Développement en série de $\ln(1-x)$

Soit x un réel de $]0,1[$.

1. Montrer, pour tout entier n de $\mathbb{N}^\times$, l'égalité : $\int\limits_0^x \dfrac{1-t^n}{1-t}\, dt = \sum\limits_{k=1}^{n} \dfrac{x^k}{k}$.

2. À l'aide d'un encadrement simple, montrer que : $\lim\limits_{n\to+\infty} \int\limits_0^x \dfrac{t^n}{1-t}dt = 0$.

3. En déduire la convergence de la série de terme général $\dfrac{x^k}{k}$ ainsi que l'égalité $\sum\limits_{k=1}^{+\infty} \dfrac{x^k}{k} = -\ln(1-x)$.

Partie III. Loi binomiale négative.

Toutes les variables aléatoires qui interviennent dans cette partie sont définies sur un même espace probabilisé $(\Omega; A; P)$.
Soit p un réel de $]0,1[$. On pose $q = 1-p$, et on considère une variable aléatoire X à valeurs dans $\mathbb{N}^\times$, qui suit la loi géométrique de paramètre p. On rappelle que pour tout entier k de $\mathbb{N}^\times$, $P\ (X = k) = pq^{k-1}$.

1. Calculer la valeur de l'espérance $E(X)$ et de la variance $V(X)$ de la variable aléatoire X.

2. On considère la variable aléatoire Y définie par $Y = \dfrac{1}{X}$.

   a) Déterminer l'ensemble des valeurs prises par Y ainsi que la loi de probabilité de Y.
   b) Montrer que Y admet une espérance $E(Y)$, que l'on calculera en fonction de p et q.
   c) Pour tout entier i supérieur ou égal à 2, établir l'existence du moment $E(Y^i)$ d'ordre i de Y.

3. Soit X_1 et X_2 deux variables aléatoires indépendantes, à valeurs dans $\mathbb{N}^\times$, qui suivent la même loi géométrique de paramètre p. On pose : $S_1 = X_1$, $S_2 = X_1 + X_2$, $Y_2 = \dfrac{2}{S_2}$.

   a) Déterminer la loi de probabilité de chacune des variables aléatoires S_2 et Y_2.
   b) Établir l'existence de l'espérance $E(Y_2)$ de la variable aléatoire Y_2.
   c) Calculer cette espérance en fonction de p et q.

4. On considère une suite $(X_n)_{n\in\mathbb{N}^\times}$ de variables aléatoires à valeurs dans $\mathbb{N}^\times$, indépendantes, de même loi géométrique de paramètre p. Pour tout entier n de $\mathbb{N}^\times$, on pose $S_n = \sum_{k=1}^{n} X_k$.

   a) Calculer l'espérance $E(S_n)$ et la variance $V(S_n)$ de la variable aléatoire S_n.
   b) Montrer que la loi de probabilité de la variable aléatoire S_n est donnée, pour tout entier s de $\mathbb{N}^\times$par :
      - si $s < n$, $P(S_n = s) = 0$
      - si $s \geqslant n$, $P(S_n = s) = \binom{s-1}{n-1} p^n q^{s-n}$.

5. Pour tout entier n de $\mathbb{N}^\times$, on pose $Y_n = \dfrac{n}{S_n}$.

   a) Préciser l'ensemble des valeurs prises par la variable aléatoire Y_n ainsi que la loi de probabilité de Y_n.
   b) Soit t un réel quelconque de $[0, 1[$. Montrer que, pour tout m de $\mathbb{Z}$, la série de terme général $s^m t^s$, $\left(\sum_{s\geqslant 1} s^m t^s\right)$, est convergente. En déduire, en particulier, l'existence des moments d'ordre 1 et 2, $E(Y_n)$ et $E(Y_n^2)$ de la variable aléatoire Y_n.

Partie IV. Une estimation ponctuelle du paramètre p.

Soit p un réel de $]0, 1[$. Dans cette partie, on considère une variable aléatoire réelle X à valeurs dans $\mathbb{N}^\times$, qui suit une loi géométrique de paramètre p inconnu. On pose $q = 1 - p$.
Pour tout entier naturel non nul n, on considère un n-échantillon $(X_1, X_2, ..., X_n)$ de variables aléatoires à valeurs dans $\mathbb{N}^\times$, indépendantes, de même loi que X.
Les variables aléatoires X; X_1; X_2;....; X_n sont définies sur un même espace probabilisé $(\Omega; A; P)$.
On pose $\overline{X}_n = \dfrac{S_n}{n} = \dfrac{1}{Y_n}$.

1. Montrer que $\overline{X}_n$ est un estimateur sans biais pour le paramètre $\dfrac{1}{p}$.

   Quel est le risque quadratique de $\overline{X}_n$ en $\dfrac{1}{p}$?

2. Pour tout entier n de $\mathbb{N}^\times$, on note h_n et φ_n les applications définies sur $[0, 1[$ à valeurs dans $\mathbb{R}$ telles que :

$$\forall t \in [0, 1[, \quad h_n(t) = \sum_{s=n}^{+\infty} \frac{1}{s}\binom{s-1}{n-1} t^s \quad , \quad \varphi_n(t) = \sum_{s=n}^{+\infty} \frac{1}{s^2}\binom{s-1}{n-1} t^s$$

On admet dans toute la suite du problème, que $h_n$est de classe C^1 et que pour tout réel t de $[0,1[$, la dérivée h_n' de $h_n$vérifie : $h_n'(t) = \sum\limits_{s=n}^{+\infty} \binom{s-1}{n-1} t^{s-1}$.

On admet également que la fonction $\varphi_n$est dérivable sur $]0;\,1[$, de dérivée φ_n', et que pour tout t de $]0,1[$, $\varphi_n'(t) = \dfrac{1}{t} h_n(t)$.

a) Montrer que $E(Y_n) = n \left(\dfrac{p}{q}\right)^n h_n(q)$.

Établir que, pour tout q de $[0,1[$, on a : $h_n(q) = \displaystyle\int_0^q \frac{t^{n-1}}{(1-t)^n} dt$.

b) À l'aide du changement de variable $y = \dfrac{t}{1-t}$, que l'on justifiera, montrer que pour tout q de $[0,1[$, $h_n(q) = \displaystyle\int_0^{q/p} \frac{y^{n-1}}{1+y} dy$.

En déduire, en utilisant une intégration par parties, que l'on peut écrire pour tout q de $[0,1[$, $h_n(q) = \dfrac{q^n}{n\,p^{n-1}} + \dfrac{1}{n} \displaystyle\int_0^{q/p} \frac{y^n}{(1+y)^2} dy$.

3. Pour tout entier n de $\mathbb{N}^\times$, soit $b_n$la fonction définie sur $]0,1[$ à valeurs réelles qui, à tout réel p de $]0,1[$ associe $b_n(p) = E(Y_n) - p$. ($b_n$représente le biais de Y_n pour estimer p)

a) Montrer que $b_n(\,p) = \left(\dfrac{p}{q}\right)^n \displaystyle\int_0^{q/p} \frac{y^n}{(1+y)^2} dy$.

b) En déduire que la suite $(b_n(p))_{n\in\mathbb{N}^\times}$ est convergente et préciser sa limite. Calculer $\lim\limits_{n\to+\infty} E(Y_n)$.

c) À l'aide d'une intégration par parties, montrer l'égalité : $b_n(p) = \dfrac{pq}{n} + o\left(\dfrac{1}{n}\right)$.

Partie V. Limite de la variance de Y_n.

Le contexte de cette partie est identique à celui de la partie IV.

1. a) En utilisant la formule établie dans la question IV.2.b, montrer que, pour tout entier n de $\mathbb{N}^\times$, $h_n(t) \sim \dfrac{t^n}{n}$ lorsque t tend vers 0 par valeurs supérieures.

b) En déduire que pour tout réel q de $]0,1[$, l'intégrale $\int\limits_0^q \dfrac{h_n\,(t)}{t} dt$ est convergente et que $\varphi_n(q) = \int\limits_0^q \dfrac{h_n\,(t)}{t} dt$.

2. a) Pour tout entier n de $\mathbb{N}^\times$, et pour tout réel t de $]0;\,1[$, on pose $H_n(t) = \int\limits_0^{t/(1-t)} \dfrac{y^n}{(1+y)^2} dy$.

Montrer que, pour tout t de $]0,1[$, on a $\varphi_n'(t) = \dfrac{1}{n}\left(\dfrac{t^{n-1}}{(1-t)^{n-1}} + \dfrac{H_n(t)}{t}\right)$

b) Établir l'existence de l'intégrale $\int\limits_0^q \dfrac{H_n(t)}{t} dt$, et en déduire l'égalité $\varphi_n(q) = \dfrac{1}{n}\left(\int\limits_0^q \dfrac{t^{n-1}}{(1-t)^{n-1}}\, dt + \int\limits_0^q \dfrac{H_n(t)}{t}\, dt\right)$

3. a) Établir l'encadrement suivant : $0 \leqslant \int\limits_0^q \dfrac{H_n(t)}{t} dt \leqslant \dfrac{h_{n+1}(q)}{n+1}$. En déduire $\lim\limits_{n\to+\infty} n\left(\dfrac{p}{q}\right)^n \int\limits_0^q \dfrac{H_n(t)}{t} dt = 0$.

b) Montrer, pour tout réel q de $]0;1[$, l'égalité $\int_0^q \left(\frac{t}{1-t}\right)^{n-1} dt = \frac{q^n}{np^{n-2}} + \frac{2}{n}\int_0^{q/p} \frac{y^n}{(1+y)^3} dy.$

En déduire que $\lim_{n\to+\infty} n\left(\frac{p}{q}\right)^n \int_0^q \left(\frac{t}{1-t}\right)^{n-1} dt = p^2$

c) On désigne par $V(Y_n)$ la variance de Y_n. Calculer $\lim_{n\to+\infty} V(Y_n)$.

Partie VI. Un intervalle de confiance du paramètre p

Dans cette partie, le contexte est identique à celui des deux parties précédentes.

1. a) En utilisant le résultat de la question IV.3.c, montrer que, lorsque n tend vers $+\infty$:

$$(E(Y_n))^2 = p^2 + \frac{2p^2\,q}{n} + o\left(\frac{1}{n}\right).$$

b) On admet que, lorsque n tend vers $+\infty$: $E(Y_n^2) = p^2 + \frac{3p^2q}{n} + o\left(\frac{1}{n}\right)$

Établir que, lorsque n tend vers $+\infty$, $V(Y_n) \sim \frac{p^2q}{n}$.

2. Pour tout entier n de $\mathbb{N}^\times$, on pose $T_n = \frac{Y_n - p}{\sqrt{\frac{p^2q}{n}}}$.

On admet que la suite de variables aléatoires $(T_n)_{n\in\mathbb{N}^\times}$ converge en loi vers une variable aléatoire T qui suit la loi normale centrée, réduite.
Cette question a pour objectif la détermination, pour n assez grand, d'un intervalle de confiance du paramètre inconnu p, au risque α donné. Autrement dit, il s'agit de trouver des variables aléatoires $I_n$et J_n, fonctions de $Y_n$telles que $P(I_n \leqslant p \leqslant J_n) = 1-\alpha$.

a) Soit a_α, le réel strictement positif tel que $P(T \geqslant a_\alpha) = \frac{\alpha}{2}$. Pour n assez grand, on peut considérer que : $P(-a_\alpha \leqslant T_n \leqslant a_\alpha) = 1-\alpha$.
En déduire l'égalité : $P\left(Y_n - a_\alpha p\sqrt{\frac{q}{n}} \leqslant p \leqslant Y_n + a_\alpha p\sqrt{\frac{q}{n}}\right) = 1-\alpha$.

b) Montrer que l'on peut choisir les “ statistiques ” $I_n$et $J_n$de la façon suivante :

$$I_n = Y_n - \frac{2a_\alpha}{3\sqrt{3}}\frac{1}{\sqrt{n}} \quad \text{et} \quad J_n = Y_n + \frac{2a_\alpha}{3\sqrt{3}}\frac{1}{\sqrt{n}}.$$

c) On suppose que $n = 900$. Un échantillon observé x_1; x_2;....; x_{900} de réalisations des variables aléatoires X_1; X_2;....; X_{900} a fourni le résultat suivant : $\overline{x}_{900} = \frac{1}{900}\sum_{i=1}^{900} x_i = 4$.
Calculer la réalisation y_{900} de la variable aléatoire Y_{900}.
On se donne un niveau de risque $\alpha = 0{,}05$; le nombre $a_{0,05}$ est à peu près égal à 2.
Sachant que $\frac{2}{45\sqrt{3}} \approx 0{,}026$, trouver un intervalle de confiance réalisé qui contienne le paramètre inconnu p avec un niveau de confiance au moins égal à 0,95.