Révisions 9

Sujet ESSEC

Sujet HEC 2005 E2

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 $1 \leq r \leq n-1$, on rappelle la formule du triangle de Pascal :

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}.$$

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

2. Soit (n, r) un couple d'entiers naturels, tels que $1 \leq r \leq n$.

   Pour tout réel $x \in]0, 1[$, on définit la fonction $f_{r,n}$ par : $\displaystyle f_{r,n}(x) = \sum_{k=r}^{n} \binom{k}{r} x^k$.

   (a) Montrer, pour tout réel $x \in]0, 1[$, que :

   $$(1-x) f_{r,n}(x) = x f_{r-1,n-1}(x) - \binom{n}{r} x^{n+1}.$$

   (b) On suppose l'entier r fixé.

   Montrer que, lorsque n tend vers $+\infty$: $\displaystyle\binom{n}{r} \underset{n \to +\infty}{\sim} \frac{n^r}{r!}$.

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 $\displaystyle\lim_{n \to +\infty} f_{0,n}(x)$ et $\displaystyle\lim_{n \to +\infty} f_{1,n}(x)$.

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

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

   Montrer que, pour tout réel $x \in]0, 1[$, $\displaystyle\lim_{n \to +\infty} f_{r,n}(x) = \frac{x^r}{(1-x)^{r+1}}$.

   (c) Conclure que, pour tout $r \in \mathbb{N}$, $\displaystyle\sum_{k=r}^{+\infty} \binom{k}{r} x^k = \frac{x^r}{(1-x)^{r+1}}$.

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

Soit x un réel de $]0, 1[$.

4. Montrer, pour tout entier $n \in \mathbb{N}^*$, l'égalité : $\displaystyle\int_0^x \frac{1-t^n}{1-t} dt = \sum_{k=1}^{n} \frac{x^k}{k}$.

5. À l'aide d'un encadrement simple, montrer que : $\lim_{n\to+\infty}\int_0^x \frac{t^n}{1-t}dt = 0$.

6. En déduire la convergence de la série de terme général $\frac{x^k}{k}$ ainsi que l'égalité : $\sum_{k=1}^{+\infty}\frac{x^k}{k} = -\ln(1-x)$.

Partie III. Loi binomiale négative.

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

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

8. On considère la variable aléatoire Y définie par $Y = \frac{1}{X}$.
   (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 $i \geq 2$, établir l'existence du moment $E(Y^i)$ d'ordre i de Y.

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

10. On considère une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires à valeurs dans $\mathbb{N}^*$, indépendantes et de même loi géométrique de paramètre p. Pour tout entier $n \in \mathbb{N}^*$, 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 $s \in \mathbb{N}^*$ par :
       - si $s < n$, $P(S_n = s) = 0$,
       - si $s \geq n$, $P(S_n = s) = \binom{s-1}{n-1} p^n q^{s-n}$.

11. Pour tout entier $n \in \mathbb{N}^*$, on pose : $Y_n = \frac{n}{S_n}$.
    (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 $m \in \mathbb{Z}$, la série $\sum_{s\geq 1} s^m t^s$ est convergente.
    En déduire l'existence des moments d'ordre 1 et 2 de la variable aléatoire Y_n.

Partie IV. Une estimation ponctuelle du paramètre p.

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

12. Montrer que $\overline{X}_n$ est un estimateur sans biais pour le paramètre $\dfrac{1}{p}$ (c'est-à-dire tel que $E(\overline{X}_n) = \dfrac{1}{p}$).

    Quel est la variance de $\overline{X}_n$?

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

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

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

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

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

    Établir que, pour tout $q \in [0,1[$, 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 $q \in [0,1[$, $h_n(q) = \displaystyle\int_0^{q/p} \frac{y^{n-1}}{1+y} dy$.

    En déduire, en utilisant une intégration par parties, que l'on peut écrire pour tout $q \in [0,1[$,

$$h_n(q) = \frac{q^n}{n\,p^{n-1}} + \frac{1}{n}\int_0^{q/p} \frac{y^n}{(1+y)^2} dy.$$

14. Pour tout entier $n \in \mathbb{N}^*$, soit b_n la fonction définie sur $]0,1[$ à valeurs réelles qui, à tout réel $p \in]0,1[$ 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 $(b_n(p))_{n\in\mathbb{N}^*}$ est convergente et préciser sa limite.
    Calculer $\displaystyle\lim_{n\to+\infty} E(Y_n)$.

    (c) À l'aide d'une intégration par parties, montrer l'égalité :

$$b_n(p) = \frac{pq}{n} + o\left(\frac{1}{n}\right).$$

Partie V. Un intervalle de confiance du paramètre p.

Dans cette partie, le contexte est identique à la partie précédente

15. (a) En utilisant le résultat de la question 14.(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 $+\infty$: $E(Y_n^2) = p^2 + \dfrac{3p^2q}{n} + o\left(\dfrac{1}{n}\right)$.

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

16. Pour tout entier $n \in \mathbb{N}^*$, on pose : $T_n = \dfrac{Y_n - p}{\sqrt{\dfrac{p^2q}{n}}}$.

On admet que la suite de variables aléatoires $(T_n)_{n\in\mathbb{N}^*}$ 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 $P(I_n \leq p \leq J_n) = 1 - \alpha$.

(a) Soit a_α le réel strictement positif tel que $P(T \geq a_\alpha) = \dfrac{\alpha}{2}$.
Pour n assez grand, on peut considérer que : $P(-a_\alpha \leq T_n \leq a_\alpha) = 1 - \alpha$.
En déduire l'égalité :

$$P\left(Y_n - a_\alpha p\sqrt{\frac{q}{n}} \leq p \leq Y_n + a_\alpha p\sqrt{\frac{q}{n}}\right) = 1 - \alpha.$$

(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 $x_1, x_2, \ldots, x_{900}$ de réalisations des variables aléatoires $X_1, X_2, \ldots, X_{900}$ a fourni le résultat suivant : $\overline{x}_{900} = \dfrac{1}{900}\displaystyle\sum_{i=1}^{900} x_i = 4$.

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 $\dfrac{2}{45\sqrt{3}} \approx 0,026$, trouver un intervalle de confiance réalisé qui contienne le paramètre inconnu p avec un niveau de confiance au moins égal à $0,95$.