Introduction à la Modélisation Statistique

Examen final – 5 Janvier 2016

Durée : 1h30

Documents interdits – calculatrices UPPA autorisées

Chaque réponse devra être justifiée et rédigée de manière rigoureuse. La qualité de la rédaction, la clarté et la précision des raisonnements interviendront pour une part importante dans l'appréciation des copies.

Tous les résultats demandés seront arrondis au 10^{-3} près.

Exercice 1 (5 pts). Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes suivant une loi Bernoulli $\mathcal{B}(p)$ de paramètre $p \in]0,1[$. On définit la v.a.d. $S_n = X_1 + \ldots + X_n$.

a) Quelle loi de probabilité suit S_n ? Justifier votre réponse.

b) Soit $\hat{p}_n = \frac{S_n}{n}$. Prouver que :

i) $\hat{p}_n$ est un estimateur sans biais du paramètre p, c.-à.-d. $\mathrm{E}[\hat{p}_n] = p$,

ii) et que $\mathrm{Var}[\hat{p}_n] \underset{n\to\infty}{\longrightarrow} 0$.

Exercice 2 (8 pts). On a créé un lac artificiel dans la côté française des Pyrénées dans lequel on a introduit des truites et des brochets en quantité égale. Au bout d'un an, on souhaite savoir quelle est la proportion p de truites parmi les poissons.
On décide de pêcher des poissons en différents endroits du lac, à différents moments, en les relâchant à chaque fois après avoir noté leur espèce. Sur 123 poissons ainsi répertoriés pendant la journée, on compte 59 truites.

a) Quel intérêt peut avoir une telle méthodologie pour la pêche, sachant que l'on veut utiliser la méthode de l'intervalle de confiance ?

b) A l'aide d'un intervalle de confiance à 95%, procéder à une estimation de la proportion de truites.

c) Peut-on écarter l'hypothèse qu'il y ait autant de brochets que de truites ? Pourquoi ?

d) Combien aurait-il fallu pêcher de poissons pour assurer un encadrement de p à 0,1 près ?

e) Au même moment, on a créé un autre lac artificiel dans la côté espagnole des Pyrénées, où on a aussi introduit des truites et des brochets en quantité égale. On reproduit notre méthodologie sur ce lac au bout d'un an, et on compte 159 truites sur 237 poissons. Peut-on assurer, avec un niveau de confiance de 95%, que les truites se sont mieux développées sous le soleil de la côté espagnole ? Justifier votre réponse.

Exercice 3 (7 pts). On dispose de deux urnes $\mathcal{U}_1$ et $\mathcal{U}_2$ contenant chacune 5 boules. Dans l'urne $\mathcal{U}_1$ il y a deux boules blanches et trois boules noires, alors que $\mathcal{U}_2$ contient une boule blanche et quatre boules noires. Un jeu consiste à lancer un dé équilibré à 6 faces : si le résultat est 2, le joueur prend une boule de $\mathcal{U}_1$, sinon il prend une boule de $\mathcal{U}_2$. Le joueur gagne s'il prend une boule blanche.
Soient les événements $D = \{\text{“ Le joueur obtient un 2 ”}\}$ et $G = \{\text{“ Le joueur gagne le jeu ”}\}$.

a) Représenter le jeu à l'aide d'un arbre pondéré en termes de D et G, en calculant les probabilités de chaque branche de l'arbre.

b) Calculer la probabilité de que le joueur gagne.

c) Un joueur a gagné ! Quelle est la probabilité que la boule provienne de l'urne $\mathcal{U}_1$?

d) Un casino propose ce jeu à ses clients, mais il soupçonne que certains clients trichent à l'heure de jouer. Le contrôle de la brigade des jeux établit que sur 1500 parties, on a compté 330 gagnants.

   i) Dire pourquoi la méthode par intervalle de fluctuation peut être utilisée dans ce cas.

   ii) Donner l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de jeux gagnés. Faire de même avec un intervalle de fluctuation asymptotique à 99%.

   iii) Est-ce que le casino a des raisons pour s'inquiéter ?

Introduction à la Modélisation Statistique

Solutions détaillées

Solution 1.

a) Chaque X_i correspond à une épreuve de Bernoulli d'un certain succès S de probabilité p, c.-à-d.

$$X_i = \begin{cases} 1 & , \text{ si on obtient un succès dans la } i\text{-ème preuve} \\ 0 & , \text{ sinon} \end{cases}$$

Alors, $S_n = X_1 + \cdots + X_n$ correspond à compter le nombre de succès parmi n épreuves de Bernoulli indépendantes de paramètre p. D'où, par définition, S_n suit une loi binomiale de paramètres n et p, c.-à-d. $S_n \sim \mathcal{B}(n, p)$.

b) On définit $\hat{p}_n = S_n/n$. D'abord, on a vu dans le cours que l'espérance et la variance d'une v.a. discrète X vérifient les propriétés suivantes :

$$\forall a, b \in \mathbb{R} : \quad \mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b \quad \text{et} \quad \mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X]. \qquad (*)$$

i) On a prouvé dans (a) que $S_n \sim \mathcal{B}(n, p)$, et on a vu dans le cours que l'espérance d'une binomiale peut s'exprimer comme $\mathrm{E}[S_n] = np$. D'où, en utilisant les conditions de linéarité $(*)$:

$$\mathrm{E}[\hat{p}_n] = \mathrm{E}\left[\frac{S_n}{n}\right] \overset{(*)}{=} \frac{1}{n}\,\mathrm{E}[S_n] = \frac{1}{n} np = p.$$

ii) De façon analogue, la variance de $S_n \sim \mathcal{B}(n, p)$ peut s'exprimer comme $\mathrm{Var}[S_n] = np(1-p)$. Par les conditions de linéarité $(*)$:

$$\mathrm{Var}[\hat{p}_n] = \mathrm{Var}\left[\frac{S_n}{n}\right] \overset{(*)}{=} \frac{1}{n^2}\,\mathrm{E}[S_n] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Solution 2. La proportion de truites parmi les poissons dans le lac est de p, donc la probabilité de $S = \{\text{“ Pecher une truite ”}\}$ est de $P(S) = p$.

a) Pour un échantillon de n poissons, on peut définir les variables aléatoires $X_1, \ldots, X_n$ tels que

$$X_i = \begin{cases} 1 & , \text{ si le } i\text{-ème poisson pêché est une truite} \\ 0 & , \text{ sinon} \end{cases}$$

En suivant cette méthodologie, on peut assumer que :
- Chaque X_i correspond à une épreuve de Bernoulli de probabilité p, car on relâche les poissons une fois leur espèce est notée, donc on ne change pas la proportion de truites.
- Les v.a. $X_1, \ldots, X_n$ sont indépendantes, car on pêche les poissons en différents endroits du lac et à différent moments.

b) Pour notre échantillon de $n = 123$ poissons, on obtient une fréquence relative de $\hat{p}_{123} = \frac{59}{123} \simeq 0,479$, d'où on construit un intervalle de confiance à 95% I_{123}. Tout d'abord, on vérifie les conditions pratiques d'approximation du Thm. de Moivre-Laplace :

$$n = 123 \geq 30, \quad n\hat{p}_n = 59 \geq 5, \quad n(1-\hat{p}_n) = 64 \geq 5.$$

On calcule l'intervalle de confiance :

$$I_{123} = \left[\hat{p}_{123} \pm 1,96\sqrt{\frac{\hat{p}_{123}(1-\hat{p}_{123})}{123}}\right] = \left[\frac{59}{123} \pm 1,96\sqrt{\frac{\frac{59}{123}\cdot\frac{64}{123}}{123}}\right] = [0,390; 0,567],$$

qui vérifie $P(p \in I_{123}) \simeq 0,95$, c.-à-d. $I_{123} = [0,390; 0,567]$ contient la vraie valeur de p dans le 95% des cas.

c) Avec une confiance du 95%, la valeur 0,5 (qui représente "l'equiproportionalité" des truites et des brochets) est contenue par l'intervalle de confiance $I_{123} = [0,390; 0,567]$. Donc on ne peut pas écarter que l'hypothèse que la vraie proportion de truites soit $p = 0,5$, avec une confiance de 95%.

d) Pour un échantillon de n poissons, on construit l'intervalle de confiance à 95% pour la vraie valeur de p :

$$I_n = \left[\hat{p}_n \pm 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right] \quad \text{tel que} \quad P(p \in I_n) \simeq 0,95.$$

Si on veut avoir un encadrement de la vraie valeur de p à $0,1$ près, la longueur de l'intervalle I_n doit être plus petite que $0,1$:

$$\text{long}(I_n) = \sup(I_n) - \inf(I_n) = 2 \times 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq 0,1.$$

La fréquence relative $\hat{p}_n$ ne peut pas être estimé que à partir des observations de l'échantillon. D'après le cours, on sait qu'on peut majorer l'expression $\hat{p}_n(1-\hat{p}_n)$ par $1/4$. C.-à-d. :

$$\text{long}(I_n) = 2 \times 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq 2 \times 1,96\sqrt{\frac{1/4}{n}} = \frac{1,96}{\sqrt{n}}.$$

De cette façon, on cherche un n minimal tel que :

$$\text{long}(I_n) \leq \frac{1,96}{\sqrt{n}} \leq 0,1,$$

soit vérifiée. On obtient :

$$\frac{1,96}{\sqrt{n}} \leq 0,1 \Longleftrightarrow n \geq \left(\frac{1,96}{0,1}\right)^2 = 384,16.$$

Donc, on a aurait besoin de pêcher au moins 358 poissons pour assurer un encadrement de la vraie valeur de p au $0,1$ près.

e) On considère que la proportion de truites parmi les poissons dans le lac de la coté espagnole est de p^{esp}, donc la probabilité de $S^{\text{esp}} = \{\text{“ Pecher une truite dans le lac espagnol ”}\}$ est de $P(S^{\text{esp}}) = p^{\text{esp}}$. De façon analogue au cas français, chaque poisson pêché correspond à une épreuve de Bernoulli :

$$Y_i = \begin{cases} 1 & \text{, si le } i\text{-ème poisson pêché est une truite} \\ 0 & \text{, sinon} \end{cases}$$

de paramètre p^{esp}. On a un échantillon de $n = 237$ poissons et fréquence relative $\hat{p}^{\text{esp}}_{237} = \frac{159}{237} \simeq 0,670$. Le conditions pratiques :

$$n = 237 \geq 30, \quad n\hat{p}^{\text{esp}}_n = 159 \geq 5, \quad n(1-\hat{p}^{\text{esp}}_n) = 78 \geq 5,$$

sont vérifiées et on peut construire l'intervalle de confiance I_{237}^{esp} à 95% pour p^{esp} :

$$I_{237} = \left[\hat{p}_{237}^{\text{esp}} \pm 1,96\sqrt{\frac{\hat{p}_{237}^{\text{esp}}(1-\hat{p}_{237}^{\text{esp}})}{237}}\right] = \left[\frac{159}{237} \pm 1,96\sqrt{\frac{\frac{159}{237}\cdot\frac{78}{237}}{237}}\right] = [0,610; 0,0,729],$$

qui vérifie $P(p^{\text{esp}} \in I_{237}^{\text{esp}}) \simeq 0,95$. On remarque que $I_{123} \cap I_{237}^{\text{esp}} = \emptyset$, en particulier :

$$\sup(I_{123}) = 0,567 < 0,610 = \inf(I_{237}^{\text{esp}}),$$

donc on peut assumer que $p < p^{\text{esp}}$ avec une confiance de 95% et que les truites de la coté espagnole se sont mieux développées que celles de la coté française.

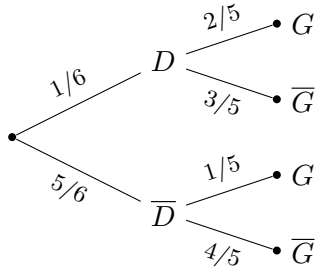
Solution 3.

a) On peut supposer, dans le lancé du dé et dans les tirages de boules des urnes, qu'on se trouve dans le modèle équiprobable. En utilisant la Règle de Laplace, on peut calculer les probabilités :

$$P(D) = \frac{1}{6}, \qquad P(\overline{D}) = \frac{5}{6},$$

$$P(\{\text{“ Obtenir blanche de l'urne } \mathcal{U}_1 \text{ ”}\}) = \frac{2}{5}, \quad P(\{\text{“ Obtenir blanche de l'urne } \mathcal{U}_2 \text{ ”}\}) = \frac{1}{5}$$

De cette façon, on a :



b) Il est clair que G est réunion disjointe de $D \cap G$ et $\overline{D} \cap G$. À partir de l'arbre pondéré précédent :

$$\begin{aligned} P(G) &= P(D \cap G) + P(\overline{D} \cap G) = P(D) \cdot P(G|D) + P(\overline{D}) \cdot P(G|\overline{D}) \\ &= \frac{1}{6} \cdot \frac{2}{5} + \frac{5}{6} \cdot \frac{1}{5} = \frac{7}{30}. \end{aligned}$$

c) On doit calculer la probabilité de D en sachant que G est réalisé, c.-à-d. :

$$P(D|G) \overset{\text{déf.}}{=} \frac{P(D \cap G)}{P(G)} = \frac{P(D) \cdot P(G|D)}{P(G)} = \frac{(1/6) \cdot (2/5)}{7/30} = \frac{2}{7}.$$

d) Chaque test correspond à une épreuve de Bernoulli :

$$X_i = \begin{cases} 1 & , \text{ si le } i\text{-ème partie donne un gagnant} \\ 0 & , \text{ sinon} \end{cases}$$

avec probabilité p. On peut supposer que les tests sont indépendants. Sur un échantillon de 1500 parties, on obtient une fréquence relative obtenue $\hat{p}_{1500} = \frac{330}{1500} = 0,22$ des gagnes.

i) On est dans la PRISE DE DÉCISION : on assume que le jeu se déroule complètement au hasard. D'après (b), cet hypothèse correspond avec :

$$\mathcal{H} : p = \frac{7}{30} \simeq 0,233.$$

Car les conditions pratiques d'approximation du Thm. De Moivre-Laplace

$$n = 1500 \geq 30, \quad np = 350 \geq 5, \quad n(1-p) = 1150 \geq 5,$$

sont vérifiées, on peut construire un intervalle de fluctuation asymptotique au 95% $I_{1500}^{95\%}$ ou 99% $I_{1500}^{99\%}$ pour la fréquence relative $\hat{p}_{1500}$ tels que :

$$P\left(\hat{p}_{1500} \in I_{1500}^{95\%}\right) \simeq 0,95, \quad P\left(\hat{p}_{1500} \in I_{1500}^{99\%}\right) \simeq 0,99.$$

De cette façon, on pourra rejeter ou non l'hypothèse $\mathcal{H}$ en fonction de si la fréquence relative appartient finalement à l'intervalle de fluctuation o non, avec un certain niveau de confiance.

ii) Les intervalles respectives viennent données par :

$$I_{1500}^{95\%} = \left[p \pm 1,96\sqrt{\frac{p(1-p)}{1500}}\right] = \left[\frac{7}{30} \pm 1,96\sqrt{\frac{(7/30)\cdot(23/30)}{1500}}\right] = [0,212; 0,254]$$

et

$$I_{1500}^{99\%} = \left[p \pm 2,58\sqrt{\frac{p(1-p)}{1500}}\right] = \left[\frac{7}{30} \pm 2,58\sqrt{\frac{(7/30)\cdot(23/30)}{1500}}\right] = [0,205; 0,261]$$

iii) En principe, le casino n'a pas des raisons pour s'inquiéter, car la fréquence relative obtenue $\hat{p}_{1500} = \frac{330}{1500} = 0,22$ appartient aux intervalles de fluctuation asymptotiques, donc on ne peut pas rejeter l'hypothèse $\mathcal{H}$.