Contrôle : Statistique inférentielle

Patrick Tardivel
Université de Bourgogne

19/05/2022

Exercice 1. *(4 points) La proportion d'informaticiennes parmi les informaticiens est de 20%. Dans un institut d'informatique de 50 personnes il y a seulement 5 informaticiennes. On pose p la probabilité inconnue que cet institut recrute une informaticienne lors d'une embauche. Peut-on dire que la probabilité de recruter une informaticienne lors d'une embauche est significativement inférieure à 20% ?*

1. *Faire un test statistique au niveau 5% pour conclure en rédigeant soigneusement chaque étape.*
2. *Calculer la p-valeur.*
3. *Calculer la puissance de ce test lorsque $p = 0,10$.*

Indication : *Soit X une variable aléatoire de loi binomiale $\mathcal{B}(50; 0,20)$. On pourra utiliser la fonction de répartition de X donnée ci-dessous :*

k	0	1	2	3	4	5	6	7
$\mathbb{P}(X \leq k)$	0,00001	0,00019	0,00128	0,00565	0,018496	0,04802	0,10339	0,19040

Soit X une variable aléatoire de loi binomiale $\mathcal{B}(50; 0,1)$. On pourra utiliser la fonction de répartition de X donnée ci-dessous :

k	0	1	2	3	4	5	6	7
$\mathbb{P}(X \leq k)$	0,00515	0,03378	0,11172	0,25029	0,43119	0,61612	0,77022	0,87785

Exercice 2. *(4 points) Deux types de calmants A et B sont administrés à des malades ayant subi une opération. Sur 754 malades ayant absorbé le calmant A, 301 ont déclaré leur douleur atténuée. Sur 587 malades ayant absorbé le calmant B, 198 ont déclaré leur douleur atténuée. Peut-on dire que les calmants ont une efficacité significativement différente.*

1. *Faire un test statistique au niveau 5% pour conclure en rédigeant soigneusement chaque étape.*
2. *Calculer la p-valeur.*

Remarque (exercices 3 et 4) : on rappelle que pour une variable aléatoire réelle R de densité f_R et pour un entier k on a $\mathbb{E}(R^k) = \int_{\mathbb{R}} x^k f_R(x)dx$.

Exercice 3. *(6 points) Soit $X_1, \dots, X_n$ un $n-$échantillon d'une loi de densité*

$$f_\theta(x) = \frac{x}{\theta} \exp\left(-\frac{x^2}{2\theta}\right) \mathbf{1}(x > 0), \text{ avec } \theta > 0.$$

1. *Vérifier que f_θ est bien une densité.*
2. *Calculer l'estimateur $\hat{\theta}$ du maximum de vraisemblance de θ.*
3. *On pose $I_k(\theta) = \int_0^{+\infty} x^k f_\theta(x)dx$. Soit $k \in \mathbb{N}$, vérifier que $I_{k+2}(\theta) = (k+2)\theta I_k(\theta)$.*
4. *Déduire de la question précédente que $\hat{\theta}$ est un estimateur sans biais pour θ puis calculer $EQM_\theta(\hat{\theta})$.*

Exercice 4. *(6 points) La méthode de Monte-Carlo permet de calculer numériquement une intégrale. Des étudiants souhaitent évaluer cette méthode sur un exemple très simple : le calcul numérique de $A = \int_0^1 x^2 dx = 1/3$ et comparer la méthode Monte-Carlo à une seconde méthode naïve.*

Comparaison des deux méthodes :

Les deux estimateurs pour A sont définis ci-dessous :

Méthode de Monte-Carlo : *Soit* $U_1, \ldots, U_n$ *des variables aléatoires iid de loi uniforme sur* $[0,1]$. *On pose* $T_1 = \frac{1}{n}\sum_{i=1}^n U_i^2$.

Méthode naïve : *Soient les variables aléatoires suivantes permettant de modéliser le tirage uniforme de points dans* $[-1,1]^2 \subset \mathbb{R}^2$ *:*

— $X_1, \ldots, X_n$ *(pour les abscisses) des variables aléatoires iid de loi uniforme sur* $[-1,1]$

— $Y_1, \ldots, Y_n$ *(pour les ordonnées) des variables aléatoires iid de loi uniforme sur* $[-1,1]$.

On admet que $\mathbb{P}(X_i^2 \leq Y_i) = A$, *il est alors naturel d'estimer* A *via la statistique suivante :* $T_2 = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i^2 \leq Y_i)$.

1. *Est-ce que* T_1 *et* T_2 *sont des estimateurs biaisés pour* A *?*
2. *Calculer l'erreur quadratique moyenne* $EQM_A(T_1)$ *et* $EQM_A(T_2)$ *des estimateurs* T_1 *et* T_2. *Quel estimateur est le plus performant pour estimer* A *au sens de l'erreur quadratique moyenne ?*

Intervalle de confiance pour A **:**

1. *Justifiez la limite suivante :*

$$\lim_{n\to+\infty} \mathbb{P}\left(T_2 - 1.96\sqrt{\frac{2}{9n}} \leq A \leq T_2 + 1.96\sqrt{\frac{2}{9n}}\right) = 0,95.$$

   En déduire le nombre n *de tirages à effectuer pour obtenir, au niveau de confiance 95%, une estimation de* A *avec la méthode naïve ayant une précision égale à* $0,01$.

2. *Justifiez la limite suivante :*

$$\lim_{n\to+\infty} \mathbb{P}\left(T_1 - 1.96\sqrt{\frac{4}{45n}} \leq A \leq T_1 + 1.96\sqrt{\frac{4}{45n}}\right) = 0,95.$$

   En déduire le nombre n *de tirages à effectuer pour obtenir, au niveau de confiance 95%, une estimation de* A *avec la méthode de Monte-Carlo ayant une précision égale à* $0,01$.

3. *On considère le code R suivant :*

```
n=1000
U=runif(n)
Estimateur=mean(U^2)
q=qnorm(0.975)
I=c(Estimateur-q*sqrt(4/45)/sqrt(n),Estimateur+q*sqrt(4/45)/sqrt(n))
```

   dans ce code, que représentent : Estimateur, q ? Comment modifier ce code pour que I *soit un intervalle de confiance pour* A *de niveau* $0,90$ *?*