ESTIMATEURS

SIMON COSTE

Exercice 1 (estimation de moyenne)**.** Jean-Michel cherche à estimer la proportion inconnue $p \in [0,1]$ de ménages possédant une navette spatiale, puis son évolution au cours du temps. À cet effet, il réalise deux enquêtes.
Pour la première enquête, il choisit au hasard et de façon indépendante $n_1 = 135633$ foyers dans la population mondiale. On appelle S_1 la variable aléatoire correspondant au nombre de foyers possédant une navette spatiale dans ce premier échantillon.
Pour la deuxième enquête, le budget de Jean-Michel a été réduit. Il choisit donc au hasard et de façon indépendante $n_2 = 23$ foyers. Il appelle S_2 la variable aléatoire correspondant au nombre de foyers possédant une navette spatiale dans ce deuxième échantillon.

(1) Quelles sont les lois de S_1 et de S_2 ? Donner leur espérance et leur variance.
(2) On définit les variables aléatoires
$$\mathrm{F}_1 = \frac{\mathrm{S}_1}{n_1} \text{ et } \mathrm{F}_2 = \frac{\mathrm{S}_2}{n_2}$$
Calculer l'espérance et la variance de F_1 et F_2.
(3) Montrer que F_1 et F_2 sont des estimateurs sans biais de p.
(4) On pose $\mathrm{G} = \frac{1}{2}(\mathrm{F}_1 + \mathrm{F}_2)$. Calculer l'espérance de G. Que peut-on en déduire pour G ?
(5) Calculer la variance de l'estimateur G.
(6) À quelle condition G est-il un meilleur estimateur que F_1 et F_2 ?
(7) De manière générale, on s'intéresse aux estimateurs de p de la forme $u\mathrm{F}_1 + v\mathrm{F}_2$, avec $u, v \in \mathbb{R}$. Aider Jean-Michel à déterminer une condition sur u et v pour que $u\mathrm{F}_1 + v\mathrm{F}_2$ soit un estimateur sans biais de p, tout en étant de variance minimale parmi les estimateurs sans biais de cette forme.

Exercice 2 (intervalles de confiance)**.** Lors d'un sondage réalisé en 2006 aux États-Unis, on a observé que $35,2\%$ des 10000 personnes interrogées soutenaient la politique de G.W. Bush. On suppose que les observations ont été réalisées de façon indépendante, non biaisée et avec remise, et on note p la proportion de la population des États-Unis effectivement en faveur de la politique de G.W. Bush au moment du sondage.

(1) Modéliser l'expérience aléatoire réalisée.
(2) Donner un intervalle de fluctuation à 95% pour la proportion de sondés favorables à la politique de G.W. Bush. Comment cet intervalle est-il construit ?

Date: 29 septembre 2020.

(3) Donner un intervalle de confiance à 95% pour p.

(4) Commenter les hypothèses faites au début de l'exercice.

Exercice 3 (mise en situation bayésienne)**.** En vous rendant à vos oraux d'agrégation, vous êtes témoin d'une conversation entre deux passagers de l'autobus.

— Jeanne-Victorine, je dois vous faire un aveu, dit brusquement la première personne. Aujourd'hui, j'ai reçu les résultats de mon test : ce dernier prétend que je suis porteur du Sars-Cov2.

— Ventrebleu ! C'est inquiétant, répond l'autre personne. Ces nouveaux tests sont performants : le taux de faux positifs est égal à 10%, et le taux de faux négatifs est égal à 9%. Jean-Victor, il y a fort à parier que vous êtes réellement malade, et vous devriez d'ailleurs porter un masque.

— Jeanne-Victorine (l'interlocuteur adopte un ton sentencieux), la proportion de personnes infectées sur l'ensemble de la population est égale à 1%. Il y a donc très peu de chances pour que je sois malade.

Les deux personnages commencent à se battre, mais aucun ne prend le dessus. Ils se tournent alors vers vous et disent :

— Il nous semble évident que vous possédez d'impressionnantes capacités d'analyse. Pouvez-vous nous aider à calculer la probabilité pour que Jean-Victor soit effectivement porteur du Sars-Cov2 ?

Exercice 4 (estimation de maximum)**.** On cherche à estimer un paramètre $\theta > 0$ à partir d'un n-échantillon $(\mathrm{X}_1, \dots, \mathrm{X}_n)$ de loi $\mathcal{U}([0, \theta])$.

(1) Calculer $\mathbf{E}[\overline{\mathrm{X}_n}]$ et en déduire un estimateur sans biais $\hat{\theta}_n$ de θ.

(2) Donner le risque quadratique de $\hat{\theta}_n$ en tant qu'estimateur de θ.

(3) Calculer la fonction de répartition puis la densité de $\tilde{\theta}_n = \max(\mathrm{X}_1, \dots, \mathrm{X}_n)$.

(4) Donner le risque quadratique de $\tilde{\theta}_n$ en tant qu'estimateur de θ.

(5) Construire un intervalle de confiance pour θ au niveau 95%.

Exercice 5 (kurtosis)**.** La *kurtosis* d'une variable aléatoire X non constante possédant des moments jusqu'à l'ordre 4 est définie par

$$\kappa(\mathrm{X}) = \frac{\mathbf{E}[|\mathrm{X} - \mu|^4]}{\mathbf{E}[|\mathrm{X} - \mu|^2]^2}$$

où $\mu = \mathbf{E}[\mathrm{X}]$.

(1) En utilisant l'inégalité de Cauchy-Schwarz [1], montrer que $\kappa(\mathrm{X}) \geqslant 1$. Que dire d'une variable aléatoire X telle que $\kappa(\mathrm{X}) = 1$?

(2) Soit $(\mathrm{X}_0, \dots, \mathrm{X}_n)$ un n-échantillon de variables aléatoires possédant des moments d'ordre $1, 2, 3, 4$. Proposer un estimateur convergent de la kurtosis.

Exercice 6. Au moment de la proclamation des résultats d'un concours, le président du jury s'aperçoit du fait qu'il a oublié de calculer la moyenne générale des notes

1. Rappel : soient deux variables aléatoires réelles A et B. Alors, $\mathbf{E}[|\mathrm{AB}|] \leqslant \sqrt{\mathbf{E}[|\mathrm{A}|^2]\mathbf{E}[|\mathrm{B}|^2]}$, et il n'y a égalité que si $\mathrm{A} = \lambda \mathrm{B}$ pour un certain λ.

obtenues par les candidats ayant composé. Il se souvient par contre du fait que la répartition des notes pouvait être approchée par une loi normale et décide d'utiliser ses capacités de statisticien pour communiquer un intervalle dans lequel se situe la moyenne inconnue avec une bonne probabilité. Pris de court, il choisit sur la liste des candidats ayant composé huit noms au hasard et consulte la moyenne obtenue par les candidats correspondants. Il obtient les nombres suivants :

$$10.22, \quad 7.31, \quad 18.01, \quad 14.98, \quad 2.25, \quad 12.20, \quad 19.38 \quad \text{et} \quad 11.07$$

(1) Si le président se souvient du fait que l'écart-type calculé de la distribution des notes est égal à 5.01, dans quel intervalle peut-il affirmer que le paramètre se situe avec une certitude de 90% ?

(2) Qu'en est-il si le président ne se souvient pas de l'écart-type de la distribution des notes ?

(3) Supposons que le président ait particulièrement à cœur de ne pas annoncer une moyenne trop basse et accorde moins d'importance au fait d'annoncer une moyenne trop haute. Dans quel intervalle peut-il affirmer que le paramètre se situe avec une certitude de 90% en tenant compte de ces contraintes ?

Exercice 7 (variation d'un paramètre)**.** Cette semaine, $n_0 = 1074309$ personnes ont effectué un test pour savoir si elles portaient un virus précis. Parmi ces personnes, $S_0 = 66671$ étaient positives. La semaine précédente, $n_1 = 1041279$ avaient été testées et $S_1 = 56227$ étaient positives. L'augmentation du taux d'incidence est-elle significative ?