UNIVERSITÉ LUMIÈRE LYON 2
LICENCE MIASHS 3ÈME ANNÉE - PARCOURS DATA SCIENCE (IDS)

Statistique Inférentielle Avancée - TP

Exercice 1 : Simulation de Monte-Carlo

On cherche dans cet exercice à approcher l'intégrale $I = \int_0^2 e^{-\frac{x^2}{2}} dx$. Pour cela nous utilisons une méthode de Monte-Carlo. Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires uniformes sur $[0, 2]$, et soit $Y_i = e^{-\frac{X_i^2}{2}}$ pour tout $i = 1, n$.

1. Quelle est la limite, au sens de la convergence en probabilité, de $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ lorsque $n \to \infty$?
2. Utiliser ce résultat pour approcher l'intégrale I, en simulant n variables aléatoires Y_i ($n = 100, 10^4, 10^6$).
3. Répéter 100 fois ces approximations, et représenter les résultats sous la forme d'une boîte à moustache pour chacune des 3 valeurs de n utilisées. Que constatez-vous ?
4. Représenter cette fois ces résultats sous la forme d'un histogramme (pour chaque valeur de n). Avez-vous une idée de la distribution de ces résultats d'approximation ? Que vous dit le théorème centrale limite ?

Exercice 2 : Calcul de vraisemblance

1. Simuler 3 échantillons $X_1, \ldots, X_n$ gaussiens d'espérance 10 et de variance 1 (fonction `rnorm`) de taille 10, 100 et 1000.
   On oublie désormais que nous avons simulé ces échantillons à partir une loi normale, et nous allons essayer plusieurs modélisation pour cette échantillon.
2. Première hypothèse : nous supposons que l'échantillon suit une loi exponentielle. Estimer le paramètre de cette loi, et calculer la vraisemblance de l'échantillon sous cette hypothèse.
3. Faites de même pour la loi normale. Que concluez-vous ?

Exercice 3 : Estimation de densité

1. Simuler trois séries de données de tailles $n = 10$, $n = 100$ et $n = 1000$ représentant des observations i.i.d. issues d'une distribution exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda = 0.5$ et $\lambda = 1$
2. Pour chaque valeur de λ et de n, représenter graphiquement la fonction de répartition empirique et théorique (sur le même graphique)
3. Pour chaque valeur de λ et de n, représenter graphiquement sur le même graphique l'estimation non-paramétrique (fonction `density`) de la fonction de densité de la loi $\mathcal{E}(\lambda)$ en utilisant les observations simulées et la densité théorique. On utilisera des noyaux gaussiens et rectangulaire (cf. le polycopié de cours).

Exercice 4 : Puissance de test

1. Créer une matrice à $N = 100$ lignes et $n = 100$ colonnes, à l'aide de la commande `matrix`.
2. Remplir chaque ligne de la matrice par un échantillon de 100 simulations de loi normale centrée réduite.
3. Créer une fonction permettant d'effectuer le test de nullité de l'espérance. Cette fonction aura en paramètre le risque de première espèce, et retournera 0 si H_0 est rejeté, 1 sinon.
4. Au risque $\alpha = 5\%$, combien de fois parmi les 100 simulations le test a-t-il accepté H_0, rejeté ?
5. Faire de même en simulant cette fois des gaussiennes centrées en 1. En déduire une valeur expérimentale de la puissance de ce test. Tester plusieurs valeurs de n (10, 50 et 100).
6. La puissance du test de nullité de la moyenne, dans les conditions de cet exercice (distribution gaussienne et variance connue égale à 1), définie par $1 - p(\text{accepter}H_0|H_1)$, est donnée par :

$$\begin{aligned} P(\mu_1) &= 1 - P(|\bar{X}| < \frac{u_{1-\alpha/2}}{\sqrt{n}} | H_1 : \bar{X} \sim \mathcal{N}(\mu_1, \frac{1}{n})) \\ &= 1 - \Phi(u_{1-\alpha/2} - \sqrt{n}\mu_1) + \Phi(-u_{1-\alpha/2} - \sqrt{n}\mu_1) \end{aligned}$$

   Programmer cette fonction puissance.

7. Représenter $P(\mu_1)$ pour $\mu_1 \in [-2, 2]$, en superposant sur un même graphique les courbes de puissance du test pour $n = 10, 50, 100$. Quel test est le plus puissant ?
8. Dans le cas où $H_1 : \mu_1 = 1$, quelle est la puissance de chaque test. Comparer avec les valeurs expérimentales obtenues en 5.

Exercice 5 : Calcul du nombre de sujets pour atteindre une puissance de test

On réalise un essai clinique ayant pour objectif d'évaluer l'efficacité d'un médicament censé réduire le taux de cholesterol dans le sang. Pour un homme de 40 ans, le taux de cholesterol moyen est de 1.8g/l. Un médicament est jugé efficace s'il réduit ce taux de $\delta = 0.2$g/l. On considère le test $H_0 : \mu = 1.8$ contre $H_1 : \mu = 1.8 - \delta$. On suppose que l'écart-type des taux de glucose est $\sigma = 0.5$, et que les taux de glucoses sont distribués normalement.

1. Pour $n = 100$ et $\alpha = 5\%$, tracer le graphique de la puissance du test $1 - \beta$ en fonction de δ (on prendra $\delta \in \{0.1, 0.2, \ldots, 1\}$.
2. Pour $\alpha = 5\%$, $\delta = 0.2$, tracer le graphique de la puissance du test $1 - \beta$ en fonction de n.
3. Si $\alpha = 5\%$, $\delta = 0.2$, quel est le nombre de sujets nécessaires à inclure dans l'essai clinique pour que le risque de seconde espèce ne dépasse pas $\beta = 5\%$.

Exercice 6

Reprendre l'exercice 2 du TP sous R (Chapitre 5.4).

Exercice 7

Reprendre l'exercice 3 du TP sous R (Chapitre 5.4).

Exercice 8

Reprendre l'exercice 5 du TP sous R (Chapitre 5.4).

Exercice 9

Sur 10 patients choisis au hasard on observe l'évolution durant 5 jours du taux (en mg/litre sang) d'une certaine substance.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Jour 1	124	88	130	115	92	80	101	98	132	85	88	110
Jour 2	125	75	138	108	92	78	105	97	125	86	87	108
Jour 3	117	73	133	108	92	74	101	92	124	83	83	110
Jour 4	123	69	130	102	88	70	95	93	128	84	83	104
Jour 5	119	70	127	98	88	70	95	93	125	85	84	103

1. Tracer sur un même graphique les 5 fonctions de répartition empiriques ainsi que les 5 boîtes à moustaches correspondant aux 5 jours.
2. Les données observées permettent-elles de conclure à une variation significative dans le temps du taux mesuré.
3. Les données observées permettent-elles de conclure à une décroissance significative dans le temps du taux mesuré.