Introduction à la Modélisation Statistique

TD3 – Échantillonage et Estimation

Exercice 1. D'après jouer au "pile-face" avec un ami, on soupçonne qu'il nous a triché et que la pièce n'est pas équilibrée. On la lance 200 fois et on obtient 120 fois "pile". Est-il vraisemblable que la pièce soit équilibrée ?

Exercice 2. Lors d'une enquête, on considère un échantillon de 1000 personnes devant représenter l'ensemble des Français (on suppose qu'il y a 50% de femme en France). Cet échantillon est composé de 441 femmes et de 559 hommes.

a) Peut-on considérer que cette échantillon est représentatif des Français à 95% près.

b) Combien de femmes au minimum aurait dû contenir cet échantillon pour être représentatif des Français à 95% près ?

Exercice 3. Dans un casino, il a été décidé que les "'machines à sous"' doivent être réglées sur une fréquence de gain du joueur de $g = 0.06$. Trois contrôleurs vérifient le réglage des machines. Le premier a joué 50 fois et a gagné 2 fois, le second a joué 120 fois et a gagné 14 fois, le troisième a joué 400 fois et a gagné 30 fois. En utilisant des intervalles de fluctuations asymptotiques au seuil 95%, donner la décision à prendre suite à chaque contrôle (à savoir valider l'hypothèse ou la refuser).

Exercice 4 (D'après un document du groupe "Statistiques et citoyenneté", IREM Paris-Nord.)**.** Woburn est une petite ville industrielle du Massachussetts, au Nord-Est des États-Unis. Dans les années 70, la communauté locale s'émeut du grand nombre d'enfants atteints de leucémie dans certains quartiers de la ville entre 1969 et 1979, on a observé douze cas de leucémie sur un échantillon de 11 748 enfants de moins de 14 ans. À cette même époque, aux États-Unis, la proportion p de leucémies chez les enfants de moins de 14 ans est $0,00045$. On fait l'hypothèse qu'à Woburn, la proportion théorique de leucémies chez les enfants de moins de 14 ans est la même que celle des États-Unis.

a) Déterminez les intervalles de fluctuation (centrés en p) à 95% et à 99% de la fréquence de leucémies sur un échantillon aléatoire de 11 748 enfants de moins de 14 ans. Écrivez les bornes de ces intervalles avec cinq décimales.

b) Peut-on considérer, au seuil de risque de 5%, que la fréquence observée de leucémies chez les enfants de moins de 14 ans à Woburn, dans cet échantillon, est en accord avec la proportion de leucémies chez les enfants de moins de 14 ans dans la population américaine ?[1] Même question mais au seuil de 1%.

Exercice 5. Un grossiste a acheté 50 000 clés USB à un fabricant qui lui a certifié que 60% avaient une capacité de 4 Go et 40% une capacité de 2 Go. Un technicien prélève au hasard 400 clés USB parmi lesquelles 220 ont une capacité de 4 Go.

a) Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de clés de 4 Go pour un échantillon de taille 400. (On donnera trois décimales).

b) Quelle hypothèse le technicien peut-il tester par cette méthode ?

c) Le technicien doit-il alerter son patron ?

1. Une enquête conduite parle Département de la Santé Publique du Massachusetts en 1981 confirmera que les sols de Woburn ont été contaminés par des résidus de tannerie et de produits chimiques. Les industriels concernés seront condamnés et la dépollution des sites sera engagée. Du point de vue étiologique, c'est l'exposition des enfants in utérus à cette eau contaminée qui serait à l'origine des cas de leucémies observés.

Exercice 6. Un fast-food a mis en un place un jeu de grattage. Pour chaque menu acheté, le client reçoit une carte à gratter. Cette carte contient 4 cases contenant chacune un symbole (invisible avant le grattage). Le client doit gratter 2 cases, et pas une de plus. S'il découvre deux symboles identiques, alors il a gagné un sandwich, sinon il a perdu. Chaque carte contient exactement une combinaison gagnante.

a) Montrer que l'on peut modéliser ce jeu par une variable aléatoire de Bernoulli dont on déterminera le paramètre.

b) Calculer la probabilité de gagner deux fois de suite.

c) Serge est un habitué de ce fast-food. On note Y_n le nombre de sandwiches que Serge a gagné après n repas. Montrer que Y_n suit une loi binomiale de paramètres n et $1/6$.

d) On se place maintenant du point de vue du restaurant. Le gérant craint que les clients puissent deviner à l'avance sous quelles cases se trouvent les deux symboles identiques. Il décide donc de vérifier que la fréquence des gains par rapport au nombre de menus vendus sur une journée est normal. Pour cela, il utilise une méthode par intervalle de fluctuation au seuil de 95%. Le gérant relève donc que, sur une journée, il a été vendu 1713 menus, et que 261 sandwichs ont été gagnés.

 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 sandwichs gagnés sur la journée complète. Faire de même avec un intervalle de fluctuation asymptotique à 97%.

 iii) Le gérant a-t-il eu raison de s'inquiéter ?

 iv) Donner l'intervalle du nombre de sandwichs gagnés dans la journée pour que le restaurant soit dans la norme.

Exercice 7. Afin d'établir le profil statistique de certains malades d'un hôpital, on prélève au hasard et avec remise 100 dossiers médicaux. Malheureusement, on constate que 20 d'entre eux sont incomplets et donc inexploitables. Si on considère qu'il faut pouvoir exploiter 1000 dossiers, combien faudra-t-il en prélever pour que cette condition soit réalisée avec une probabilité égale à 95% ?

Exercice 8. On cherche à estimer la proportion d'étudiants à l'UPPA qui possèdent un ordinateur de la marque "Pomme". Pour cela, on prend un échantillon de n étudiants. Calculer la taille minimale n afin de garantir, avec un niveau de confiance de 95%, que l'erreur de l'estimation ne dépasse pas 2%.

Exercice 9. Une usine fabrique des pièces métalliques, qui sont censées résister à certaines contraintes mécaniques. Le responsable de la fabrication veut faire estimer le taux de pièces défectueuses. Pour cela, il utilise une méthode par intervalle de confiance au niveau 95%, en extrayant au hasard n pièces qui sont ensuite soumises aux contraintes mécaniques jusqu'à rupture de la pièce. Selon le niveau de résistance, la pièce sera jugée défectueuse ou non.

a) Chaque pièce testée étant détruite, le responsable souhaite minimiser la taille de l'échantillon à tester, tout en conservant un intervalle de confiance de longueur inférieure à 0,1. Quelle taille d'échantillon peut-on lui conseiller ?

b) Il est décidé de mener l'étude sur 500 pièces ; on en trouve 40 défectueuses. Quel intervalle de confiance, au niveau 95%, obtient-on ?

c) L'année précédente, les tests avaient montré 130 pièces défectueuses pour un échantillon de 1000 pièces. Peut-on supposer que les tests effectués cette année sont meilleurs ?

Exercice 10. Lors d'une élection on réalise un sondage "sortie des urnes" auprès de 500 électeurs choisis de manière aléatoire. 51% ont voté pour Madame Y, et 49% pour Monsieur Z.

a) Procéder à une estimation de la proportion d'habitants ayant voté pour Monsieur Z, à l'aide d'un intervalle de confiance au niveau 95%.

b) Peut-on garantir la défaite de Monsieur Z ?

c) À fréquences observées constantes, combien d'électeurs aurait-il fallu interroger pour prévoir valablement la victoire de Madame Y ?

Exercice 11. On a fait pousser 200 plants de courgettes avec nitrate et 200 sans nitrate. Au bout de 4 jours, le pourcentage de germination est de 45% avec nitrate et de 34% sans nitrate.

a) Procéder à une estimation de la proportion de germination dans chacun des cas, en utilisant un intervalle de confiance au niveau 95%.

b) Peut-on écarter à ce stade l'hypothèse que les germinations sont plus importantes avec nitrate ?

c) Quelle taille aurait dû avoir chacun des deux groupes pour pouvoir conclure sur l'efficacité du nitrate, en supposant que les fréquences observées n'aient pas varié.

Exercice 12. D'après une étude statistique, le pourcentage de notes supérieures à 16 attribuées aux élèves de première année universitaire est de 10%.

a) En 2010, dans un groupe de 300 étudiants, 40 notes supérieures à 16 ont été attribuées. Déterminer l'intervalle de fluctuation asymptotiques au seuil de 95% du pourcentage d'élèves ayant une note supérieure à 16 pour un échantillon de 300 personnes. Peut-on considérer qu'il s'agit d'un groupe exceptionnellement bon ?

b) En 2012, dans un groupe de 150 élèves, 11 notes supérieures à 16 ont été attribuées. Déterminer l'intervalle de fluctuation asymptotiques au seuil de 95% du pourcentage d'élèves ayant une note supérieure à 16 pour un échantillon de 150 personnes. Peut-on considérer qu'il s'agit d'un groupe exceptionnellement mauvais ?

c) Un responsable administratif de l'université affirme que ces résultats montrent une baisse du niveau entre 2010 et 2012. Son affirmation est-elle fondée ?

Exercice 13. Un programme permet, par simulations, d'afficher la fréquence d'apparition du 1 au terme de n lancers de dé parfait. Cependant, au préalable et sans le divulguer, ce programme choisit au hasard le nombre de face du dé simulé : soit 6, soit 8, soit 12.

Estimer la valeur minimale de n qui permet de dévoiler, avec un risque d'erreur d'au plus 5%, le nombre de faces initialement choisi par le programme.

Exercice 14. Une urne contient un nombre inconnu de jetons rouges et de jetons blancs. Un échantillon de 60 jetons, constitué par des tirages aléatoires avec remise jeton par jeton, montre une proportion de 70% de jetons rouges.

a) Déterminer l'intervalle de confiance au niveau de confiance du 95% pour la proportion de jetons rouges. Préciser l'amplitude de cet intervalle, exprimée en pourcentages.

b) De quelle taille devrait être l'échantillon pour que l'amplitude de l'intervalle de confiance au niveau de confiance de 95% soit inférieure ou égale à 3% ?

c) Le nombre de jetons dans le sac impose-t-il des contraintes sur la taille de l'échantillon ? Ce nombre a-t-il un impact sur l'intervalle de confiance ?

Exercice 15. L'entreprise Médiamétrie mesure l'audience des chaînes de télévision en France en se basant sur un échantillon constitué des membres du panel ayant indiqué leur présence devant leur téléviseur. On considère dans cet exercice que ces échantillons sont aléatoires.

a) Combien de téléspectateurs doivent avoir indiqué leur présence pour pouvoir construire un intervalle de confiance au niveau de confiance 0,95 d'amplitude inférieure ou égale à 7% ?

b) Un mardi à 14h, sur 1 435 téléspectateurs, 512 regardaient la première chaîne. Déterminer l'intervalle de confiance au niveau de confiance 0,95 de la proportion de téléspectateurs de la première chaîne à ce moment.

c) Un samedi à 21 h 30, sur 9 351 téléspectateurs, 3 337 regardaient la deuxième chaîne. Déterminer l'intervalle de confiance au niveau de confiance 0,95 de la proportion de téléspectateurs de la deuxième chaîne à ce moment.

d) Ces résultats permettent-ils de comparer les audiences des deux chaînes ?