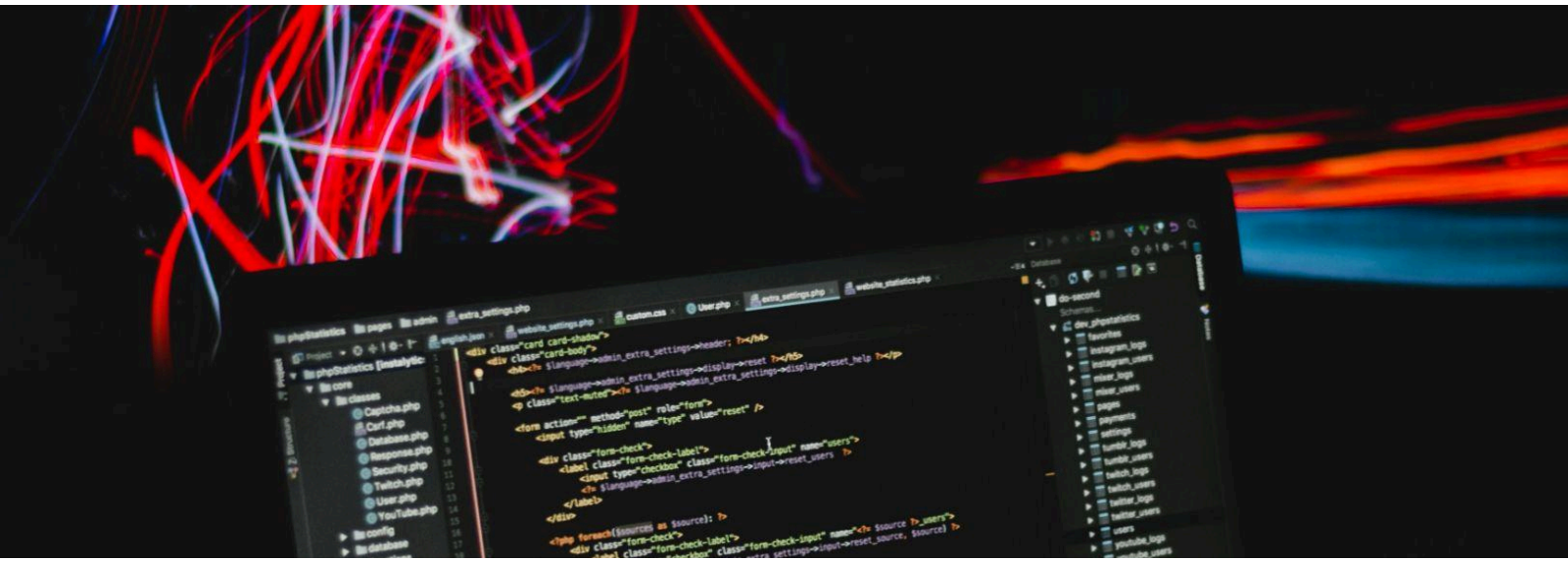
DataBird

Statistiques



L’inférence statistique

Voici une série d’exercices sur l’échantillonnage.

Exercice 1

Quelle serait la taille minimale de l'échantillon nécessaire pour obtenir une estimation précise de la qualité de production d'une entreprise de téléviseurs qui produit 1000 téléviseurs par jour, avec une marge d'erreur acceptable de 3% et un niveau de confiance de 95% ?

Exercice 2

Dans l’exercice suivant, nous allons te décrire des cas d’usages qui présentent des fonctionnements utilisés pour créer des échantillons.
Pour chaque cas, tu devras préciser s’il s’agit de l'échantillonnage probabiliste ou non, et quelle méthode a été utilisée :

1. Tu es propriétaire d’un cinéma et tu planifies d'y organiser un festival de films d’horreur le mois prochain. Pour déterminer quels films d’horreur tu présenteras, tu souhaites demander à des cinéphiles les films qu’ils préfèrent parmi une sélection de films.
   Pour dresser la liste des films nécessaires à ton sondage, tu décides d’échantillonner 10 des 100 meilleurs films d’horreur de tous les temps.
   L’une des façons d’obtenir un échantillon consisterait à écrire tous les titres des

films sur des bouts de papier, à les placer dans une boîte et à tirer les 10 titres qui constitueront ton échantillon.

2. Ton entreprise compte 100 employés. Tu souhaites connaître la satisfaction de tes employés. Pour ce faire, tu décides de faire ton étude sur 10 de tes employés.
   Tu as une liste avec les employés (classés par ordre alphabétique), tu décides d'interroger l'employé X présent sur cette liste puis d'interroger un employé toutes les 10 lignes.

3. Tu souhaites étudier les préférences musicales des étudiants d'une université. La population totale est composée de 10 000 étudiants de différentes facultés : 3 000 étudiants en sciences, 4 000 étudiants en arts et 3 000 étudiants en commerce.
   Tu divises la population en trois sous-groupes basés sur la faculté d'appartenance des étudiants. Puis, tu sélectionnes aléatoirement un échantillon dans chaque sous-groupe.

4. Tu souhaites étudier les habitudes de consommation alimentaire dans une ville. La population totale de la ville est de 50 000 habitants, répartis dans 10 quartiers différents.
   Tu sélectionnes aléatoirement la population de deux quartiers qui devient ton échantillon.

5. Tu es restaurateur et souhaites définir les habitudes de consommation de tes clients. Pour ce faire, tu as repéré une table d'une vingtaine de personnes qui te semblent représentatives du client type.
   Tu décides donc de mener ton analyse sur cet échantillon.

6. Tu es restaurateur à la fin de ton service et tu invites tes clients à te communiquer leur avis sur le service.

Exercice 3

Identifier les biais d'échantillonnage et expliquer comment ils pourraient affecter négativement les études. Puis, proposer une solution qui pourrait permettre de réduire ce biais.

1. Tu es responsable du recrutement pour une entreprise et tu dois embaucher un nouveau groupe d'employés. Pour faciliter le processus de sélection, tu organises une journée de recrutement où les candidats intéressés peuvent se présenter et participer à des entretiens d'embauche. Cependant, lors de cette journée de recrutement, tu remarques que la majorité des candidats présents

sont des étudiants de premier cycle d'une université locale. Tu reçois très peu de candidatures de personnes ayant de l'expérience professionnelle ou de candidats issus d'autres établissements d'enseignement.

2. Tu étudies la satisfaction des clients dans un restaurant. Pour collecter les données, tu distribues des questionnaires de satisfaction à la fin du repas et tu demandes aux clients de les remplir avant de partir. Cependant, tu remarques que la majorité des questionnaires sont remplis par des clients qui ont eu une expérience extrêmement positive ou extrêmement négative.

3. Tu étudies les facteurs de réussite des entrepreneurs en analysant une base de données d'entreprises. Tu collectes des données sur les caractéristiques et les performances des entreprises, telles que leur taille, leur secteur d'activité, leurs revenus, etc.
   Cependant, tu remarques que ta base de données ne contient que des entreprises existantes et en activité. Tu n’as pas d'informations sur les entreprises qui ont échoué ou qui ont été fermées par le passé.

Exercice 4

Tu es un chercheur en anthropologie et tu souhaites estimer la taille moyenne des individus d'une population donnée. Pour cela, tu réalises une étude en sélectionnant un échantillon de 100 individus. Cependant, lors de ton processus de sélection, tu commets un biais en excluant les personnes de moins de 18 ans de ton échantillon.

Les données que tu as collectées auprès des 100 individus de ton échantillon ont une moyenne de 170 cm avec un écart type de 10 cm.

1. Calculer le biais de ton estimation en comparant la taille moyenne de ton échantillon à la vraie taille moyenne de la population, qui est de 165 cm.

2. Discuter de l'impact de ce biais sur ton estimation de la taille moyenne de la population.

3. Ensuite, calculer la variance de ton échantillon.

4. Analyser l'importance de la variance dans ton estimation de la taille moyenne.

Exercice 5

Tu es un.e économiste chargé.e d'estimer le taux de chômage dans une région donnée. Tu réalises une enquête auprès d'un échantillon de 500 individus sélectionnés de manière aléatoire. Sur cet échantillon, tu constates que 40 individus sont au chômage.

1. Calculer l'estimation ponctuelle du taux de chômage dans la région à partir de ton échantillon.

2. Tu souhaites maintenant construire un intervalle de confiance à 95% pour ton estimation. Utiliser la formule de l'intervalle de confiance pour la proportion dans une population.

3. Calculer l'intervalle de confiance à 95% pour ton estimation du taux de chômage.

4. Interpréter cet intervalle de confiance

Exercice 6 (à réaliser avec Python)

Supposons que tu aies collecté un échantillon de données sur les scores de performance d'une équipe de football composée de 20 joueurs. Tu souhaites estimer la moyenne de performance de l'équipe et obtenir une mesure de l'incertitude associée à cette estimation en utilisant la méthode du bootstrapping.

Voici les scores de performance de l'équipe : 75, 80, 85, 90, 85, 90, 95, 80, 85, 90, 75, 80, 85, 90, 85, 90, 95, 80, 85, 90.

1. Effectuer une estimation ponctuelle de la moyenne de performance de l'équipe à partir de l'échantillon initial.
2. Appliquer la méthode du bootstrapping pour obtenir un échantillon bootstrap à partir de l'échantillon initial. Utiliser la méthode de l'échantillonnage avec remplacement pour créer un nouvel échantillon de la même taille que l'échantillon initial.
3. Calculer la moyenne de performance pour cet échantillon bootstrap.
4. Répéter les étapes 2 et 3 un grand nombre de fois (par exemple, 1000 fois) pour obtenir une distribution des moyennes bootstrap.
5. Utiliser cette distribution pour construire un intervalle de confiance à 95% pour la moyenne de performance de l'équipe.

Une fois que tu as réalisé ces exercices, tu peux venir télécharger le notebook d’exercices de la session.