

Examen Statistiques CPES

Durée : 2h

Documents et notes de cours interdits, calculatrice autorisée mais non utile.

Toutes les réponses doivent être justifiées pour être comptabilisées.

Le sujet est peut-être long, vous n'avez pas à le finir pour avoir une bonne note.

Laissez les résultats en fractions irréductibles.

Questions de cours:

1. Donner des exemples de variables discrètes qualitatives, quantitatives discrètes et quantitatives continues.
2. Pour les variables énoncées ci-dessus, dans quel cas utilise-t-on des modalités ou des classes pour les compartimenter ?

Soit une expérience de ayant pour résultat les données $\vec{x} = (x_1, x_2, \dots, x_n)$, obtenues par la réalisation de $\vec{X} = (X_1, X_2, \dots, X_n)$, indépendantes et identiquement distribuées.

3. Donner la formule de la moyenne empirique $\bar{x}_n$ et de la variance empirique.
4. En introduisant le vecteur réordonné $\vec{x}^*$, donner la formule de la médiane empirique. Quel peut être son intérêt lors de la description d'un jeu de données ? Donner la nature des objets mathématiques x_1, X_1 .
5. Définir la notion d'indépendance pour X_1, X_2 par exemple.
6. Que signifie *identiquement distribuées* ?
7. Soit $Var(X_1) = \mathbb{E}[(X_1 - \mathbb{E}(X_1))^2]$ la variance de X_1 . Montrer que $Var(X_1) = \mathbb{E}(X_1^2) - (\mathbb{E}(X_1))^2$.

Exercice 1

Le tableau suivant représente les poids respectifs d'un échantillon de 12 pères (X) et leurs fils aînés (Y) en kg.

I

1. Dessiner le diagramme de fréquences des deux jeux de données en regroupant les poids dans des intervalles de largeur 2 kg ($[61;63[, [63;65[, \dots]$)
2. Calculer la moyenne empirique des deux jeux de données.
3. Pour le **poids du père**, calculer sa médiane, premier et troisième quartiles, premier et neuvième déciles.
4. Dessiner le diagramme de Tukey correspondant, en renseignant toutes les quantités calculées dans la question précédente (sachant qu'il y a plusieurs conventions pour la composition de ces diagrammes, je serai tolérant sur leur construction).

X	65	63	70	61	63	67	62	66	67	80	81	64
Y	63	64	67	61	80	68	63	64	65	75	76	66

II

1. Donner la formule de la covariance $Cov(X, Y)$ et du coefficient de corrélation empirique $r(X, Y)$.
2. Supposons qu'on veuille effectuer un test d'hypothèses pour affirmer qu'il existe une corrélation entre le poids d'un père et le poids de son fils. Formuler une hypothèse nulle $\mathcal{H}_0$ et une hypothèse alternative $\mathcal{H}_1$ en fonction de $r(X, Y)$. De quel type de test s'agit-il ?
3. En rentrant les données dans un logiciel de statistiques, on obtient que pour un seuil fixé α , on ne se situe pas dans la région critique. Quelle conclusion peut-on en tirer ?

Examen Statistiques CPES

Exercice 2:

Des chercheurs ont construit un petit labyrinthe pour y faire évoluer des souris et mener certaines études. Ce labyrinthe est symétrique (par rapport à une diagonale) et comporte deux sorties, dans deux coins opposés. En préalable de leurs travaux, ils veulent déterminer la probabilité qu'une souris, déposée au centre du labyrinthe, réussisse à sortir au bout d'un temps fixé à l'avance. Pour ce faire, ils étudient la variable aléatoire X à valeurs dans $0, 1, 2$, de loi définie par:

$$\mathbb{P}(X = 0) = 1 - 2p, \mathbb{P}(X = 1) = p, \mathbb{P}(X = 2) = p, \text{ avec } p \in]0, 1/2[$$

À n reprises, ils déposent une souris (différente à chaque fois) au centre et notent le résultat. Ils obtiennent un n -échantillon $(x_1, \dots, x_n)$, à partir duquel ils souhaitent estimer p .

I.

1. Calculer l'espérance et la variance de X .
2. Rappeler le théorème de transfert pour une fonction f appliquée à la variable aléatoire X .
3. Soit $Y = X^2$, calculer l'espérance et la variance de Y .

II. Soit $X_1, X_2, \dots, X_n$ une suite de variables aléatoires indépendantes de même loi que X .

1. On considère $S_n = \sum_{i=1}^n X_i$, $T_n = \frac{S_n}{n}$. De quel paramètre T_n est-il un estimateur ? Pour S_n ?
2. Calculer $\mathbb{E}(S_n)$ et $\mathbb{E}(T_n)$. Sont-ils sans biais ?
3. Calculer $\text{Var}(S_n)$ et $\text{Var}(T_n)$, en détaillant bien les étapes de calcul.
4. Exprimer l'erreur quadratique moyenne pour un estimateur quelconque, puis donner celles de S_n et T_n .
5. Comparer ces erreurs quadratiques moyennes. Que concluez-vous quand à l'utilisation d'un estimateur par rapport à l'autre ?
6. Calculer $\lim_{n \rightarrow \infty} T_n$. Avec quel théorème ce résultat est-il en accord ?

III.

1. Soit $\theta = p$ le paramètre inconnu à estimer. Exprimer θ en fonction de $\mathbb{E}(X)$, montrer qu'il existe β tel que $T_n^{(1)} = \frac{1}{\beta} T_n$ est un estimateur sans biais de p .
2. Calculer $\mathbb{E}(X^2)$, exprimer p en fonction de $\mathbb{E}(X^2)$. Trouver une constante γ telle que $T_n^{(2)} = \frac{1}{n\gamma} \sum_{i=1}^n X_i^2$ soit un estimateur sans biais de p .
3. Calculer $\text{Var}(T_n^{(2)})$ et sa limite.

IV. Pour $n = 100$, on a observé $\sum_{i=1}^n x_i = 75$ et $\sum_{i=1}^n x_i^2 = 121$.

1. Donner deux estimations $\hat{\theta}_1 = t_n^{(1)}$ et $\hat{\theta}_2 = t_n^{(2)}$ de p , basée sur les questions précédentes. Sont-elles sans biais ?
2. En comparant les variances, quel estimateur semble le plus performant ?
3. En calculant l'espérance de $\sum_{i=1}^n (x_i - T_n)^2$, donner un estimateur sans biais de la variance.

Examen Statistiques CPES

Exercice 3:

En génétique, les feuilles sont divisées en 4 catégories différentes : amidonnée et verte, sucrée et blanche, amidonnée et blanche, et sucrée et verte. Chaque catégorie a une probabilité d'occurrence donnée par $p_1 = (\theta + 2)/4$, $p_2 = \theta/4$, $p_3 = (1 - \theta)/4$, $p_4 = (1 - \theta)/4$ respectivement, où $0 < \theta < 1$ est inconnu.

Supposons que l'on dispose de n feuilles.

Le nombre de feuilles amidonnées et vertes est modélisée par une variable N_1 de loi binomiale

$\mathcal{B}(n, p_1)$, le nombre de feuilles sucrées et blanches par N_2 de loi binomiale $\mathcal{B}(n, p_2)$.

Catégorie	Nombre
Amidonnée Verte	1997
Sucrée Blanche	32
Amidonnée Blanche	906
Sucrée Verte	904

Considérons les deux estimateurs suivants pour θ :

$$T_1 = \frac{4}{n}N_1 - 2, T_2 = \frac{4}{n}N_2$$

I

1. Montrer que T_1 et T_2 sont des estimateurs sans biais de θ .
2. Calculer leur variance.
3. Calculer l'estimation de θ donnée par chaque estimateur.

II

On pose la variable aléatoire suivante :

$$X = \begin{cases} 1 & \text{si la feuille observée est de type amidonnée verte} \\ 2 & \text{si la feuille observée est de type sucrée blanche} \\ 3 & \text{si la feuille observée est de type amidonnée blanche} \\ 4 & \text{si la feuille observée est de type sucrée verte.} \end{cases}$$

Donnée par :

a	1	2	3	4
$p(a)$	$(2 + \theta)/4$	$\theta/4$	$(1 - \theta)/4$	$(1 - \theta)/4$

1. Écrire la fonction de vraisemblance $L(x_1, \dots, x_n | \theta) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n | \theta)$ pour le problème étudié. Ici les X_i sont les réalisations de chaque expérience. Préciser les étapes où on utilise les propriétés d'indépendance et de distribution identique des variables aléatoires X_i .
2. Calculer la log-vraisemblance $l(\theta)$.
3. Donner l'expression générale de la dérivée de la log-vraisemblance $\frac{d}{d\theta}l(\theta)$.

Examen Statistiques CPES

Exercice 4:

On effectue un contrôle de fabrication sur des pièces dont une proportion p est défectueuse.

1. Pour une loi quelconque de moyenne m inconnue et de variance σ^2 connue, avec n observations, rappeler la formule générale pour les intervalles de confiance I_γ pour m , en fonction de la moyenne estimée $\hat{m}$, et de σ^2 , n et des valeurs η_γ , ξ_γ associées au niveau de confiance γ .
2. On contrôle un lot de 200 pièces et on trouve 11 pièces défectueuses. Donner la loi de probabilité adaptée à cette expérience. Calculer sa moyenne estimée, ainsi que sa variance estimée.
3. En approximant la variance par la variance estimée, donner les intervalles de confiance pour l'estimation de p , de confiance $\gamma = 0.95$. On donne η_γ , $\xi_\gamma = 1.96$.
4. Autre approche : pour $y \in]0,1[$, calculer les variations de la fonction $f(y) = y(1 - y)$ et déterminer sa borne supérieure. Déterminer un autre intervalle de confiance, qui ne dépend plus de la variance dans ce cas.
5. Dessiner les deux intervalles de confiance, en représentant leur largeur.
6. Quel est l'avantage d'utiliser le deuxième intervalle de confiance ? Quel peut être son inconvénient ?
7. Dans les deux cas, combien faudra-t-il contrôler de pièces pour avoir, au risque de $\gamma = 0.95$, un intervalle de confiance de largeur 0.2% ?
8. On souhaite effectuer un test d'hypothèses pour étudier la performance du processus de fabrication des pièces. On se demande si la proportionne pièces défectueuses dépasse une valeur p_0 fixée. Formuler une hypothèse nulle $\mathcal{H}_0$ et une hypothèse alternative $\mathcal{H}_1$ en fonction de p . De quel type de test s'agit-il ?