The National School of Artificial Intelligence ENSIA		Statistical inference Semester 4 2023/2024

Worksheet n°1

Exercise 1. The table below shows the distribution of a university's 10,000 undergraduate students by level of study and gender:

Level / Gender	S_1	S_2	S_3	S_4	S_5	S_6
Male	2000	1600	1000	900	400	250
Female	1500	900	600	500	200	150

We propose to carry out a survey of 1000 students to obtain information on the quality of teaching at this university. How can this sample be distributed among the different strata of the 1000 students, taking into account level and gender?

Le tableau ci-dessous, établi la répartition des 10 000 étudiants d'une université en licence selon le niveau d'étude et le sexe :

Niveau / Sexe	S_1	S_2	S_3	S_4	S_5	S_6
Masculin	*2000*	*1600*	*1000*	*900*	*400*	*250*
Féminin	*1500*	*900*	*600*	*500*	*200*	*150*

On se propose de réaliser une enquête auprès de 1000 étudiants pour obtenir des informations sur la qualité de l'enseignement au sein de cette université. Comment répartir cet échantillon entre les différentes strates des 1000 étudiants, en tenant compte du niveau et du sexe.

Exercise 2. On a population of 15,000 companies studied according to sales (variable X) and number of employees (variable Y). We obtain the following results:

X (10^5 DZD)	Number	Y	Number
[0,50[	5500	[0,10[	7050
[50,100[	4500	[10,50[	3700
[100,200[	1750	[50,100[	2500
[200,500[	2000	[100,500[	1500
[500,1000[	500	[500,--- [	250
[1000,--- [	750	-	-
Total	15000	Total	15000

Consider a sample of 1,500**.** What method can be used to determine the sample? Assume that the institute responsible for the survey has 30 interviewers.
Explain your approach and the concepts used.

Sur une population de 15000 entreprises étudiées selon le chiffre d'affaires (variable X) et le nombre d'employés (variable Y). On obtient les résultats suivants :

X (10^5 DZD)	*Effectif*	*Y*	*Effectif*
[0,50[	*5500*	*[0,10[*	*7050*
[50,100[	*4500*	*[10,50[*	*3700*
[100,200[	*1750*	*[50,100[*	*2500*
[200,500[	*2000*	*[100,500[*	*1500*
[500,1000[	*500*	*[500,--- [*	*250*
[1000,--- [	*750*	-	-
Total	*15000*	*Total*	*15000*

On considère un échantillon de 1500. Quelle méthode peut-on utiliser pour déterminer un échantillon ? On supposera que l'institut responsable du sondage dispose de 30 enquêteurs.
Expliquez votre démarche et les concepts utilisés.

Exercise 3. Let T_1 and T_2 be two different estimators of the same parameter θ. Assume that $E[T_1]= \theta+b_1$ and $E[T_2]= \theta+b_2$ (b_1 and b_2 are two known numerical values). Let $T=\alpha T_1+\beta T_2$.
Determine α and β so that T is an unbiased estimator of θ:

a) When $b_1 \neq b_2$.
b) When $b_1 = b_2$.

Soient T_1 et T_2 deux estimateurs différents de même paramètre θ. On suppose que $E[T_1]= \theta+b_1$ et $E[T_2]= \theta+b_2$ (b_1 et b_2 sont deux valeurs numériques connues). Soit $T=\alpha T_1+\beta T_2$.
Déterminer α et β pour que T soit un estimateur sans biais de θ :

a) Lorsque $b_1 \neq b_2$.
b) Lorsque $b_1 = b_2$.

Exercise 4. The elements of a population possess a character X that follows a density distribution

$$f_\theta(x) = \frac{2}{\sqrt{\pi}\theta^{3/2}} x^2 e^{-x^2/\theta}.$$

Where $\theta > 0$. A series of n independent experiments yielded the values $x_1, x_2, \dots, x_n$.
1. Determine an estimator $\hat{\theta}$ of the parameter θ using the maximum likelihood method.
2. Is this estimator unbiased? Convergent? Efficient?

Les éléments d'une population possèdent un caractère X qui suit une loi de densité

$$f_\theta(x) = \frac{2}{\sqrt{\pi}\theta^{3/2}} x^2 e^{-x^2/\theta}.$$

Où $\theta > 0$. Une suite de n expériences indépendantes a donné les valeurs $x_1, x_2, \dots, x_n$.

1. *Déterminer un estimateur $\hat{\theta}$ du paramètre θ par la méthode du maximum de vraisemblance.*
2. *Cet estimateur est-il sans biais ? Convergent ? Efficace ?*

Exercise 5. In order to improve customer satisfaction, a major ISP compiles statistics on the number of hotline calls received, so as to evaluate the waiting time for the customer and the number of employees needed to staff the switchboard. The results of the survey cover 200 consecutive one-minute sequences, during which the average number of calls was 3 per minute. Calls are assumed to be evenly distributed over time: a time interval is divided into one-second units, so in each time unit there is at most one call.

1. What is the probability distribution of the number of calls received in 4 minutes?
2. Show that this distribution can be approximated by a Poisson distribution.
3. Give a confidence interval for the average number of calls in 4 minutes.

Afin de mieux satisfaire leurs clients, une grande société fournisseur d'accès internet fait ses statistiques sur le nombre d'appels reçus en hotline, elle pourra ainsi évaluer le temps d'attente pour le client et le nombre d'employés à mettre au standard; les résultats de l'enquête portent sur 200 séquences consécutives de une minute, durant lesquelles le nombre d'appels moyen a été de 3 appels par minute. On suppose que *les appels sont répartis également dans le temps : on partage un intervalle de temps en unités de une seconde; alors dans chaque unité de temps, il y a au plus un appel.*

1. *Quelle est la loi de probabilité du nombre d'appels reçus en 4 minutes ?*
2. *Montrer que l'on peut approcher cette loi par une loi de Poisson.*
3. *Donner un intervalle de confiance pour le nombre moyen d'appels en 4 minutes.*

Exercise 6. The medical staff of a large company compile statistics on the cholesterol levels of their employees; the observations on 100 employees selected at random are as follows.

Cholestérol level in cg : (center of classes)	Number of employes
120	9
160	22
200	25
240	21
280	16
320	7

1. Calculate the mean m_e and standard deviation σ_e on the sample.
2. Estimate the mean and standard deviation for cholesterol levels throughout the company.
3. Determine a confidence interval for the mean.
4. Determine the minimum sample size so that the amplitude of the confidence interval is less than 10.

Le staff médical d'une grande entreprise fait ses petites statistiques sur le taux de cholestérol de ses employés; les observations sur 100 employés tirés au sort sont les suivantes.

Taux de cholestérol en cg :(centre classe)	*Effectif d'employés*
120	*9*
160	*22*
200	*25*
240	*21*
280	*16*
320	*7*

1. *Calculer la moyenne m_e et l'écart-type σ_e sur l'échantillon.*
2. *Estimer la moyenne et l'écart-type pour le taux de cholestérol dans toute l'entreprise.*
3. *Déterminer un intervalle de confiance pour la moyenne.*
4. *Déterminer la taille minimum d'échantillon pour que l'amplitude de l'intervalle de confiance soit inférieure à 10.*

Exercise 7. The average height of a random sample of 40 people taken from a population of 780 is 1.70m. The standard deviation for the whole population is 24cm. Find the 95% confidence interval for the mean height of the population.

La taille moyenne d'un échantillon aléatoire de 40 personnes extrait d'une population de 780 individus est de 1,70m. L'écart-type pour toute la population vaut 24cm. Trouver l'intervalle de confiance pour la taille moyenne de la population à 95%.

Exercise 8. 500 students sit an exam. A random sample of 38 marks gives a mean equal to 8.65 and a standard deviation equal to 2.82. Find the confidence interval for the population mean scores at 90%, 95% and 99%.

500 étudiants se présentent à un examen. Un échantillon aléatoire de 38 notes donne une moyenne égale à 8,65 et un écart-type égal à 2,82. Trouver l'intervalle de confiance pour la moyenne des notes de la population à 90%,95% et 99%.

Exercise 9. It is assumed that the weights of 3000 students at a university follow a normal distribution with mean 68kg and standard deviation σ=3kg.

1. What is the sampling mean and standard deviation of the means if we extract 80 samples of 25 students each.
   a) In the case of a non-exhaustive draw.
   b) In the case of exhaustive sampling.
2. For how many samples can we expect to find an average:
   a) Between 68.3kg and 68.8kg.
   b) Less than 68.4kg.

On suppose que les poids de 3000 étudiants d'une université suivent une loi normale de moyenne 68kg et d'écart type σ=3kg.

1. *Quelle est la moyenne et l'écart-type d'échantillonnage des moyennes si l'on extrait 80 échantillons de 25 étudiants chacun.*
   a) Dans le cas d'un tirage non exhaustif.
   b) Dans le cas d'un tirage exhaustif
2. *Pour combien d'échantillons peut-on s'attendre à trouver une moyenne :*
   a) Comprise entre 68,3kg et 68,8kg.
   b) Inférieure à 68,4kg.

Exercise 10. The human resources manager of a company has established that the results of a test measuring the manual dexterity of the workforce assigned to tasks involving the assembly of complex parts are distributed according to the normal distribution of mean m = 72 and variance σ^2= 36.

1. What is the probability that a randomly selected employee will score less than 63 on the manual dexterity test?
2. A random sample of 25 employees took the manual dexterity test.

a) What is the distribution of the sample mean?
b) What are the mean and standard deviation of the distribution of the mean?

3. What is the probability that the sample mean lies between 69 and 75?
4. What is the probability that the standard deviation between the sample mean and the population mean is greater than 3?

Le directeur de ressources humaines d'une entreprise a établi que les résultats à un test mesurant la dextérité manuelle de la main d'œuvre affectée à des taches d'assemblages de pièces complexes sont distribués d'après la loi normale de moyenne m = 72 et de variance σ^2= 36.

1. *Quelle est la probabilité qu'un employé sélectionné au hasard obtienne un résultat inférieur à 63 au test de dextérité manuelle?*
2. *Un échantillon aléatoire de 25 employés a subi le test de dextérité manuelle.*
   a) Quelle est la distribution de la moyenne de l'échantillon?
   b) Quels sont la moyenne et l'écart-type de la distribution de la moyenne ?
3. *Quelle est la probabilité que la moyenne de cet échantillon se situe entre 69 et 75 ?*
4. *Quelle est la probabilité que l'écart-type entre la moyenne de cet échantillon et celle de la population soit supérieur à 3 ?*

Exercise 11. It is assumed that students in a statistics lecture have normally distributed grades with mean m = 72 and standard deviation σ= 9.

1. Find the probability that a single randomly selected student has a grade above 80.

2. Find the probability that a random sample of 10 students will have an average grade above 80.
3. Answer the previous question (2.) assuming that the population does not follow a normal distribution.

On suppose que les étudiants d'un cours de statistique aient des notes normalement distribuées avec une moyenne $m = 72$ et un écart-type $\sigma = 9$.

1. *Trouver la probabilité pour qu'un seul étudiant choisit au hasard ait une note supérieure à 80.*
2. *Trouver la probabilité pour qu'un échantillon aléatoire de 10 étudiants ait une note moyenne supérieure à 80.*
3. *Répondre à la question précédente (2.) en supposant que la population ne suit pas une loi normale.*

Exercise 12. Two samples of size $n_1 = 120$ and $n_2 = 150$ were taken independently from a population of economics students.
We find that 48 students in the first sample and 66 students in the second sample have a scientific secondary education. Let p be the proportion of students in the population with a scientific background.
Calculate three point estimates of p.

Dans une population d'étudiants en économie, on a prélevé, indépendamment, deux échantillon de taille $n_1 = 120$ et $n_2 = 150$.
On constate que 48 étudiants du premier échantillon et 66 étudiants du deuxième échantillon ont une formation secondaire scientifique. Soit p la proportion d'étudiants de la population ayant une formation scientifique.
Calculer trois estimations ponctuelles de p.

Exercise 13. A company uses insulating material in the assembly of a certain type of electric motor. It is important not only that the average thickness of the components meets the company's requirements, but also that the variability in thickness does not fluctuate too widely.
A random sample of 20 insulating components taken from a batch gives the following thicknesses:

Thickness in mm									
5,6	5,9	6,2	6,1	6,6	5,9	5,9	5,6	6,2	5,8
5,5	5,6	6	6,3	6,2	5,9	6,2	6	6,2	6,3

1. Assuming that the thickness of this insulating material is distributed according to a normal distribution, estimate by confidence interval the standard deviation of the thickness for the entire production run. Use a confidence level of 95%.
2. Estimate also the mean thickness of the insulating material for the whole production with a confidence level of 95%.

Une entreprise utilise une matière isolante dans l'assemblage d'un certain type de moteurs électriques. Il est important que non seulement l'épaisseur moyenne des composants rencontre les exigences de l'entreprise mais également que la variabilité de l'épaisseur ne présente pas de trop fortes fluctuations.
Un échantillon aléatoire de 20 composantes isolantes prélevé d'un lot donne les épaisseurs suivantes :

Epaisseur en mm									
5,6	*5,9*	*6,2*	*6,1*	*6,6*	*5,9*	*5,9*	*5,6*	*6,2*	*5,8*
5,5	*5,6*	*6*	*6,3*	*6,2*	*5,9*	*6,2*	*6*	*6,2*	*6,3*

1. *En admettant que l'épaisseur de cette matière isolante est distribuée selon une loi normale, estimer par intervalle de confiance l'écart-type de l'épaisseur pour l'ensemble de la production. Utiliser un niveau de confiance de 95%.*
2. *Estimer également l'épaisseur moyenne de la matière isolante pour l'ensemble de la production avec un niveau de confiance de 95%.*