

Examen statistiques CPES 1

Durée : 2h

Documents, notes de cours et calculatrice interdits.

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.

Au sein d'un même exercice, les parties A, B sont indépendantes.

Exercice 1 :

Soit X la surface d'une maison mesurée en m^2 . Le traitement de l'information relatif à 100 maisons a permis de dresser l'histogramme de la variable statistique X .

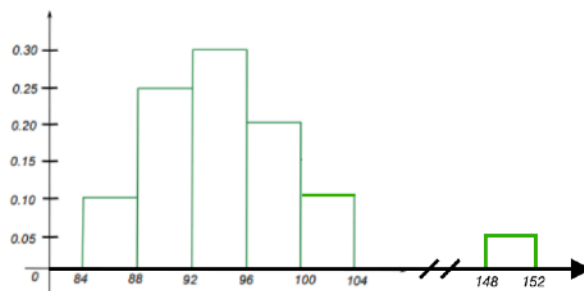


FIGURE 1 – Histogramme de la surface des maisons

L'axe des abscisses correspond à la surface des maisons en m^2 , celui des ordonnées à la proportion de foyers associée.

1. À partir de l'histogramme, comment calculeriez-vous la moyenne empirique de l'échantillon ?
On donne sa valeur $\bar{x} = 96.6 m^2$.
2. Définir la médiane empirique $\tilde{x}$ de l'échantillon et la déterminer à partir de l'histogramme.
3. Faire de même pour le premier et le troisième quartile, $q_{1/4}$, $q_{3/4}$.
4. Dessiner le diagramme de Tukey (ou boîte à moustaches) associé aux jeux de données.
5. Justifier quant à la pertinence de la médiane par rapport à la moyenne empirique.
6. Ces données sont utilisées dans le cadre d'une étude sur les consommations liées au chauffage de foyers en France. Quelle critique pourriez-vous apporter sur les données ?
7. Si on suppose que la surface d'une maison peut être modélisée par une loi normale $\mathcal{N}(\mu, \sigma)$, avec $\mu \approx \bar{x}$ et $\sigma = 5m^2$. En se rappelant les encadrements pour la loi normale, quelle information nous fournit σ ?

Exercice 2 :

On dispose d'un échantillon de données $\vec{x} = x_1, \dots, x_n \in \mathbb{R}^+$. Ces données, recueillies auprès de n foyers, sont des données de durée de vie d'un grille pain de la marque *Stroumpf* et il est commun de modéliser ce type de phénomène de durée de vie par une loi de Weibull. On les modélise donc par un ensemble de variables aléatoires $X_1, \dots, X_n$, i.i.d. provenant de la loi de Weibull à deux paramètres $\theta = (\beta, \eta)$, dont la densité est, pour $\beta, \eta > 0$:

$$\forall x \in \mathbb{R}^+, f(x|\theta) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\eta}\right)^\beta\right)$$

1. En partant des valeurs de $x \rightarrow 0^+$ et $x \rightarrow +\infty$, tracer approximativement les lois pour $\beta = 1$, puis pour $\beta = 2$.
2. En TD, pour laquelle de ces deux valeurs de β a-t-on rencontré une loi avec une forme analogue ? Donner, si possible, son nom.
3. Avec des termes « simples », comment définiriez-vous la fonction de vraisemblance ?
4. Rappeler les étapes de l'estimation par maximum de vraisemblance.
5. Écrire la fonction de vraisemblance $\mathcal{L}(\theta)$ pour ce modèle.

On fait maintenant un changement de variable en écrivant notre loi de Weibull à deux

paramètres avec $\theta' = (\beta, \gamma)$ ainsi : $\forall x \in \mathbb{R}^+, f(x|\theta') = \frac{\beta}{\gamma} x^{\beta-1} \exp\left(-\frac{x^\beta}{\gamma}\right)$.

6. Expliquer comment sont liés les vecteurs θ et θ' .
7. Écrire la log-vraisemblance $l(\theta')$ et calculer les dérivées par rapport à chacun des paramètres : $\frac{d}{d\beta} l(\theta)$ et $\frac{d}{d\gamma} l(\theta)$. Mais ne pas chercher à les résoudre. Pourquoi passe-t-on par le logarithme de la vraisemblance ?
8. On suppose β connu, fixé. Trouver l'estimateur maximum de vraisemblance pour le paramètre γ par rapport à nos données $x_1, \dots, x_n$.

Exercice 3:

Soient $n \in \mathbb{N}^*$ et $(x_1, \dots, x_n)$ un échantillon pouvant être observé par un ensemble de variables aléatoires $X_1, \dots, X_n$ suivant la loi de Poisson $\mathcal{P}(\lambda)$, $\lambda > 0$. La loi est définie de manière suivante:

$$\forall k \in \mathbb{N}, \mathbb{P}[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}$$

Le paramètre λ est à estimer.

A/ Performance de deux estimateurs

1. Définir pour le paramètre λ la fonction de vraisemblance $\mathcal{L}(\lambda)$, avec l'échantillon de données $(x_1, \dots, x_n)$.
2. Écrire la fonction de log-vraisemblance $l(\lambda) = \ln \mathcal{L}(\lambda)$ pour $(x_1, \dots, x_n)$.
3. Trouver le maximum de $l(\lambda)$ qu'on notera λ^* .
4. Avec des termes simples, qu'est-ce que décrivent le biais et la variance d'un estimateur ? Vous pouvez vous appuyer sur un schéma.
5. Calculer $\mathbb{E}[\lambda^*]$. λ^* est-il sans biais ?
6. Calculer $V(\lambda^*)$ et sa limite lorsque $n \rightarrow \infty$. Interpréter le résultat.
7. Calculer l'erreur quadratique moyenne de λ^* , $EQM(\lambda^*)$.
8. Soit un nouvel estimateur : $\hat{\lambda} = \frac{1}{3}(2X_1 + X_2)$. Calculer $\mathbb{E}[\hat{\lambda}]$. $\hat{\lambda}$ est-il sans biais ?
9. Calculer $V(\hat{\lambda})$ puis $EQM(\hat{\lambda})$. Interpréter le résultat.
10. À partir de quelle valeur de n a-t-on $EQM(\lambda^*) < EQM(\hat{\lambda})$?

B/ Estimation ensembliste de λ

On souhaite construire un intervalle de confiance pour λ à partir de l'estimation ponctuelle donnée par notre EMV λ^* . On suppose que $n = 350$ et que $\lambda^* = 3.50$.

1. Écrire la formule de l'intervalle de confiance pour le paramètre λ , en approximant l'écart-type σ^2 par l'écart-type empirique S_X^2 . On prendra soin de justifier quelles valeurs sont prises pour la moyenne empirique $\bar{X}$ et la variance empirique S_X^2 pour la loi de Poisson.
2. Pour un seuil de confiance $\gamma = 95\%$, calculer cet intervalle.
3. On considère le test d'hypothèse suivant : $\mathcal{H}_0 : \lambda = \lambda_0$ contre $\mathcal{H}_1 : \lambda \neq \lambda_0$. Comment s'appelle $\mathcal{H}_0$? Et $\mathcal{H}_1$? De quel type de test s'agit-il ?

4. La probabilité de rejeter $\mathcal{H}_0$ à tort s'écrit $\mathbb{P}[|\tilde{X}| > \eta_\gamma | \mathcal{H}_0] = \alpha = 1 - \gamma$, avec $\tilde{X} = \frac{\bar{X} - \lambda_0}{\sqrt{\frac{S_X^2}{n}}}$

Quel type d'erreur est associé à cette probabilité ?

Exercice 4 :

A/ Intervalle de confiance: construction théorique

On dispose d'un ensemble de n variables aléatoires $X_1, \dots, X_n$, i.i.d. de loi inconnue. On cherche un intervalle de confiance pour l'espérance $\mathbb{E}[X_i] = \mu$. On suppose la variance $V(X_i) = \sigma^2$ connue.

1. Donner l'expression de l'intervalle de confiance I_γ de seuil γ pour l'espérance μ . Le dessiner, en indiquant le centre de cet intervalle et sa largeur.

On se propose ensuite de redémontrer cette formule.

2. Écrire la fonction pivotale $\tilde{X}$ pour les $X_1, \dots, X_n$ en fonction de $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, μ , σ et n .

3. Calculer l'espérance et la variance de $\tilde{X}$.

4. Énoncer le célèbre théorème utilisé en statistiques qui permet de déduire la loi de $\tilde{X}$ lorsque $n \rightarrow \infty$. En pratique, à partir de quelle valeur de n ce théorème est-il applicable ?

5. On se donne une valeur de notre seuil de confiance $\gamma \in [0,1]$. On peut trouver un réel η_γ tel que $\mathbb{P}[|\tilde{X}| < \eta_\gamma] = \gamma$. En partant de $|\tilde{X}| < \eta_\gamma$, déterminer, avec un raisonnement par équivalence, un encadrement de μ en fonction des quantités définies dans la question 1 et de η_γ .

B/ Etude quantitative

Une société de vente à distance de matériel informatique s'intéresse au nombre journalier de connexions sur son site internet. Sur une période de $n = 25$ jours, les nombres suivants ont été relevés :

759 750 755 ... 767

On suppose que ces résultats sont indépendants et identiquement distribués selon une loi normale d'espérance μ et de variance σ^2

1. Le nombre journalier de connexions a été choisi comme indicateur de l'attractivité du site. Proposez une critique de se baser sur cette quantité, et suggérer une autre grandeur et/ou une modification de l'expérience.

On souhaite déterminer un intervalle de confiance pour l'espérance μ . On suppose que les X_i suivent une loi normale $\mathcal{N}(\mu, \sigma)$. La moyenne estimée est $\bar{X} = 760$, et on suppose tout d'abord que la variance est connue : $\sigma^2 = 49$.

2. Quelle loi suit la fonction pivotale $\tilde{X} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$? Donner son espérance et sa variance.

3. Donner un intervalle de confiance pour $\gamma = 95\%$.

4. Sur combien de jours aurait-on dû relever le nombre de connexions pour que la largeur totale de l'intervalle de confiance, de niveau de 95%, n'excède pas 2 (en supposant que les estimations des moyennes et variances ne changent pas).

5. Dessiner les deux intervalles de confiance.

6. Supposons maintenant que la variance soit inconnue. Est-ce plus une situation plus réaliste que dans la question 2 ? Quelle loi suit alors la fonction pivotale $Z = \frac{\bar{X} - \mu}{\sqrt{\frac{S_X^2}{n}}}$?

7. La variance estimée est $S_X^2 \approx 36$. Donner un intervalle de confiance pour $\gamma = 95\%$. On donne $t_{n-1, 1-\gamma} = t_{24, 0.05} \approx 2.5$