EXAMEN STATISTIQUE - 2ÈME ANNÉE APPRENTISSAGE

Lundi 17 Novembre 2025 (14h-15h30)

Partiel avec documents autorisés

Exercice 1 : Estimation (10 points)

On considère n observations $x_1, ..., x_n$ issues d'un vecteur de n variables aléatoires X_i indépendantes à valeurs dans $\{-1, +1\}$ telles que

$$P[X_i = 1] = p \text{ et } P[X_i = -1] = q = 1 - p$$

avec $p \in]0, 1[$ un paramètre inconnu.

1. (2pts) Montrer que

$$P[X_i = x_i] = p^{\frac{1+x_i}{2}} (1-p)^{\frac{1-x_i}{2}}, \quad \forall x_i \in \{-1, +1\}.$$

On admet que $X_i = 2Y_i - 1$ où Y_i est de loi de Bernoulli $Be(p)$ de paramètre p. En déduire que $E[X_i] = 2p - 1$ et $\text{var}[X_i] = 4pq$.

En remplaçant x_i par les valeurs 1 et -1 dans l'expression ci-dessus, on obtient

$$P[X_i = 1] = p \text{ et } P[X_i = -1] = 1 - p = q,$$

ce qui est en accord avec la définition de ces probabilités données en début d'exercice. En utilisant les propriétés de l'espérance et de la variance et en utilisant les tables pour avoir $E[Y_i] = p$ et $\text{var}[X_i] = pq$, on obtient

$$E[X_i] = 2E[Y_i] - 1 = 2p - 1 \text{ et } \text{var}[Y_i] = 4\text{var}[X_i] = 4pq.$$

2. (2pts) Déterminer la vraisemblance de $(x_1, ..., x_n)$ et montrer qu'elle s'écrit

$$L(x_1, ..., x_n; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \text{ avec } y_i = \frac{1+x_i}{2}.$$

En déduire que l'estimateur du maximum de vraisemblance du paramètre p noté $\widehat{p}_{\text{MV}}$ est défini par

$$\widehat{p}_{\text{MV}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i + 1\right).$$

On effectue les traitements habituels

- Calcul de la vraisemblance.

$$L(x_1, ..., x_n; p) = \prod_{i=1}^{n} P[X_i = x_i; p] = \prod_{i=1}^{n} p^{\frac{1+x_i}{2}} (1-p)^{\frac{1-x_i}{2}},$$

que l'on peut exprimer en fonction des variables y_i :

$$L(x_1, ..., x_n; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}, \text{ avec } y_i = \frac{1+x_i}{2}.$$

- Calcul de la log vraisemblance

$$\ln L(x_1, ..., x_n; p) = \sum_{i=1}^{n} y_i \ln(p) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - p)$$

- Calcul de la dérivée de la log-vraisemblance et de l'estimateur

$$\frac{\partial \ln L(x_1, ..., x_n; p)}{\partial p} = \sum_{i=1}^{n} y_i \frac{1}{p} - \sum_{i=1}^{n} (1 - y_i) \frac{1}{1 - p}$$

qui s'annule si

$$\frac{1}{p} \sum_{i=1}^{n} y_i - \frac{1}{1 - p} \left(n - \sum_{i=1}^{n} y_i \right) = 0$$

En multipliant cette équation par $p(1 - p)$, on obtient

$$(1 - p) \sum_{i=1}^{n} y_i - p \left(n - \sum_{i=1}^{n} y_i \right) = 0 \Leftrightarrow \sum_{i=1}^{n} y_i = np$$

On en déduit que l'estimateur du maximum de vraisemblance du paramètre p noté $\widehat{p}_{\text{MV}}$ est défini par

$$\widehat{p}_{\text{MV}} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^{n} X_i + 1 \right).$$

3. (2pts) L'estimateur $\widehat{p}_{\text{MV}}$ est-il sans biais et convergent ?
On a

$$E\left(\widehat{p}_{\text{MV}}\right) = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] = p.$$

$\widehat{p}_{\text{MV}}$ est donc un estimateur non biaisé de p. La variance de cet estimateur est (en utilisant l'indépendance entre les variables Y_i)

$$\text{var}\left(\widehat{p}_{\text{MV}}\right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}(Y_i) = \frac{pq}{n} = \frac{p(1 - p)}{n}.$$

Comme $\widehat{p}_{\text{MV}}$ est un estimateur non biaisé de p et que sa variance tend vers 0 lorsque $n \to \infty$, $\widehat{p}_{\text{MV}}$ est un estimateur convergent de p.

4. (2pts) Déterminer la borne de Cramer-Rao pour un estimateur non biaisé du paramètre p. L'estimateur $\widehat{p}_{\text{MV}}$ est-il l'estimateur efficace du paramètre p ?
La dérivée seconde de la log-vraisemblance est

$$\frac{\partial^2 \ln L(x_1, ..., x_n; p)}{\partial p^2} = -\frac{1}{p^2} \sum_{i=1}^{n} y_i - \frac{1}{(1 - p)^2} \left(n - \sum_{i=1}^{n} y_i \right),$$

d'où

$$E\left[-\frac{\partial^2 \ln L(X_1, ..., X_n; p)}{\partial p^2} \right] = -\frac{1}{p^2} \sum_{i=1}^{n} E[Y_i] - \frac{1}{(1 - p)^2} \left(n - \sum_{i=1}^{n} E[Y_i] \right).$$

Comme $E[Y_i] = p$, on a

$$E\left[-\frac{\partial^2 \ln L(X_1, ..., X_n; p)}{\partial p^2} \right] = -\frac{1}{p^2}(np) - \frac{1}{(1 - p)^2}[n(1 - p)] = \frac{-n}{p(1 - p)}.$$

On en déduit que la borne de Cramér-Rao pour un estimateur non-biaisé de p est

$$\text{BCR} = \frac{-1}{E\left[\frac{\partial^2 \ln L(X_1,...,X_n;p)}{\partial p^2}\right]} = \frac{p(1-p)}{n}.$$

Comme $\text{var}\left[\widehat{p}_{\text{MV}}\right] = \text{BCR}$ et que l'estimateur $\widehat{p}_{\text{MV}}$ est non biaisé, $\widehat{p}_{\text{MV}}$ est l'estimateur efficace du paramètre p.

5. (1pt) Déterminer l'estimateur des moments de p défini à partir de $E[X_i]$.
   D'après la première question
   $$E[X_i] = 2p - 1.$$
   On en déduit
   $$p = \frac{1 + E[X_i]}{2}$$
   et donc
   $$\widehat{p}_{\text{Mo}} = \frac{1 + \frac{1}{n}\sum_{i=1}^n X_i}{2} = \widehat{p}_{\text{MV}}.$$

6. (1pt) Quel est l'estimateur du maximum de vraisemblance du paramètre $\theta = 2p - 1$?
   En utilisant la propriété d'invariance fonctionnelle, on en obtient
   $$\widehat{\theta}_{\text{MV}} = 2\widehat{p}_{\text{MV}} - 1 = \frac{1}{n}\sum_{i=1}^n X_i.$$

Exercice 2 : Tests Statistiques (10 points)

On considère n observations $x_1, ..., x_n$ issues d'un vecteur $(X_1, ..., X_n)$ de n variables aléatoires indépendantes de même densité :

$$p(x_i;\theta) = \begin{cases} \frac{\theta}{x_i^{\theta+1}} \text{ si } x_i > 1 \\ 0 \text{ sinon} \end{cases}$$

avec $\theta > 0$. On désire utiliser les observations $x_1, ..., x_n$ pour déterminer si $\theta = \theta_0 > 0$ ou si $\theta = \theta_1 > \theta_0$. On considère donc le test d'hypothèses

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1 \qquad \text{avec } \theta_1 > \theta_0 > 0.$$

1. (2pts) Montrer que la statistique du test de Neyman Pearson est $T_n = \sum_{i=1}^n \ln(X_i)$ et que la région critique associée est $\{(x_1, ..., x_n) \in \mathbb{R}^n | T_n < S_\alpha\}$.
   Le test de Neyman Pearson est défini par
   $$\text{Rejet de } H_0 \text{ si } \frac{L(x_1, ..., x_n;\theta_1)}{L(x_1, ..., x_n;\theta_0)} > S_{1,\alpha}$$
   où $S_{1,\alpha}$ est un seuil dépendant du risque de première espèce α. Mais
   $$\begin{aligned}
   \frac{L(x_1, ..., x_n;\theta_1)}{L(x_1, ..., x_n;\theta_0)} > S_{1,\alpha} &\Leftrightarrow \ln\left[\frac{L(x_1, ..., x_n;\theta_1)}{L(x_1, ..., x_n;\theta_0)}\right] > S_{2,\alpha} \\
   &\Leftrightarrow \ln\left[\frac{\prod_{i=1}^n \frac{\theta_1}{x_i^{\theta_1+1}}}{\prod_{i=1}^n \frac{\theta_0}{x_i^{\theta_0+1}}}\right] > S_{2,\alpha} \\
   &\Leftrightarrow [(\theta_0+1) - (\theta_1+1)]\sum_{i=1}^n \ln(x_i) > S_{3,\alpha} \\
   &\Leftrightarrow (\theta_0 - \theta_1)\sum_{i=1}^n \ln(x_i) > S_{3,\alpha}.
   \end{aligned}$$

Comme $\theta_1 > \theta_0$, on rejette H_0 si

$$T_n = \sum_{i=1}^{n} \ln(X_i) < S_\alpha.$$

La région critique du test est donc l'ensemble des vecteurs $(x_1, ..., x_n) \in \mathbb{R}^n$ tels que $T_n < S_\alpha$ et la statistique de test est $T_n = \sum_{i=1}^{n} \ln(X_i)$.

2. (2pts) Montrer que $Y_i = \ln(X_i)$ possède une loi gamma $\mathcal{G}(1, \theta)$. En déduire sans calcul la moyenne et la variance de Y_i.
La variable aléatoire Y_i est à valeurs dans $\mathbb{R}^+$. En faisant un changement de variables (et en faisant attention de ne pas oublier le Jacobien ;-)), on obtient la densité de Y_i

$$g(y_i; \theta) = \begin{cases} \theta \exp(-\theta y_i) & \text{si } y_i > 0 \\ 0 & \text{sinon} \end{cases}$$

On reconnaît une loi gamma $\mathcal{G}(1, \theta)$ (ou loi exponentielle) dont la moyenne et la variance (données dans la table) sont

$$E[Y_i] = \frac{1}{\theta} \text{ et } \operatorname{var}[Y_i] = \frac{1}{\theta^2}.$$

3. (2pts) On suppose que pour n grand, on peut approcher la loi de T_n par une loi normale. Quels sont les paramètres de cette loi normale ?
Comme $T_n = \sum_{i=1}^{n} \ln(X_i) = \sum_{i=1}^{n} Y_i$, on a

$$E[T_n] = \sum_{i=1}^{n} E[Y_i] = \sum_{i=1}^{n} \frac{1}{\theta} = \frac{n}{\theta}$$

et (comme les variables Y_i sont indépendantes)

$$\operatorname{var}[T_n] = \sum_{i=1}^{n} \operatorname{var}[Y_i] = \sum_{i=1}^{n} \left(\frac{1}{\theta^2}\right) = \frac{n}{\theta^2}$$

La loi approchée de T_n est donc la loi normale $\mathcal{N}\left(\frac{n}{\theta}, \frac{n}{\theta^2}\right)$.

4. (2pts) On note F la fonction de répartition d'une loi du normale $\mathcal{N}(0, 1)$. Exprimer le risque de première espèce α en fonction du seuil du test de Neyman Pearson noté S_α, de F, n et de θ_0. En déduire la valeur de S_α en fonction de $F^{-1}(\alpha)$, n et θ_0.
Le risque α est défini par

$$\alpha = P[\text{Rejeter } H_0 | H_0 \text{ vraie}] = P\left[T_n < S_\alpha | T_n \sim \mathcal{N}\left(\frac{n}{\theta_0}, \frac{n}{\theta_0^2}\right)\right].$$

On sait que la variable $U_n = \frac{T_n - \frac{n}{\theta_0}}{\sqrt{\frac{n}{\theta_0^2}}} \sim \mathcal{N}(0, 1)$. Donc

$$\alpha = F\left[\frac{S_\alpha - \frac{n}{\theta_0}}{\sqrt{\frac{n}{\theta_0^2}}}\right],$$

d'où

$$S_\alpha = \frac{n}{\theta_0} + F^{-1}(\alpha)\sqrt{\frac{n}{\theta_0^2}}.$$

5. (2pts) Déterminer les caractéristiques opérationnelles du récepteur (COR) pour ce test en fonction de n, $F^{-1}(\alpha)$, $\frac{\theta_1}{\theta_0}$. Représenter l'allure de ces courbes COR pour différentes valeurs de n.
La puissance du test est définie par

$$\pi = P[\text{Rejeter } H_0 | H_1 \text{ vraie}].$$

On sait qu'on peut l'obtenir en remplaçant la valeur de θ_0 dans α par θ_1. On obtient alors :

$$\pi = F\left[\frac{S_\alpha - \frac{n}{\theta_1}}{\sqrt{\frac{n}{\theta_1^2}}}\right],$$

En remplaçant l'expression du seuil S_α déterminée précédemment, on obtient

$$\pi = F\left[F^{-1}(\alpha)\frac{\theta_1}{\theta_0} + \sqrt{n}\left(\frac{\theta_1}{\theta_0} - 1\right)\right].$$

L'allure des courbes COR pour différentes valeurs de n est représentée ci-dessous :

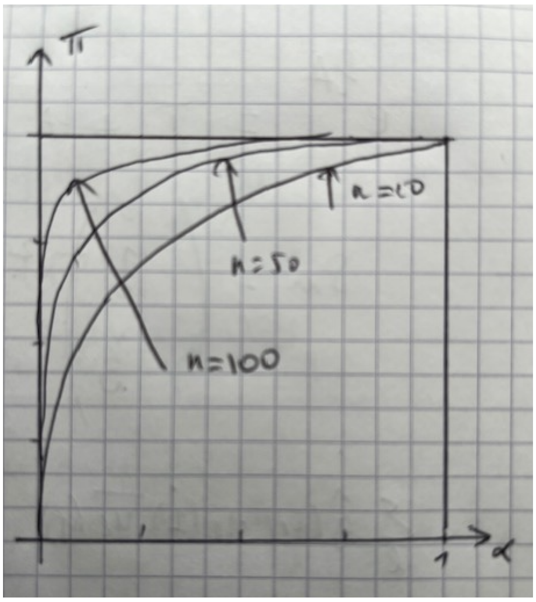

Figure 1: Allure des courbes COR pour différentes valeurs de n.

LOIS DE PROBABILITÉ CONTINUES

$\mathbf{m}$: moyenne σ^2 : variance **F. C. :** fonction caractéristique

LOI	Densité de probabilité	$\mathbf{m}$	σ^2	F. C.
Uniforme	$f(x) = \frac{1}{b-a}$ $x \in]a, b[$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\dfrac{e^{itb} - e^{ita}}{it(b-a)}$
Gamma $\mathcal{G}(\nu, \theta)$	$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\theta x} x^{\nu-1}$ $\theta > 0,\ \nu > 0$ $x \geq 0$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\dfrac{\nu}{\theta}$	$\dfrac{\nu}{\theta^2}$	$\dfrac{1}{\left(1 - i\frac{t}{\theta}\right)^\nu}$
Inverse gamma $\mathcal{IG}(\nu, \theta)$	$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\frac{\theta}{x}} \frac{1}{x^{\nu+1}}$ $\theta > 0,\ \nu > 0$ $x \geq 0$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{\theta}{\nu-1}$ si $\nu > 1$	$\frac{\theta^2}{(\nu-1)^2(\nu-2)}$ si $\nu > 2$	(*)
Première loi de Laplace	$f(x) = \frac{1}{2} e^{-\lvert x \rvert}, \quad x \in \mathbb{R}$	0	2	$\dfrac{1}{1+t^2}$
Normale univariée $\mathcal{N}\left(m, \sigma^2\right)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$	m	σ^2	$e^{imt - \frac{\sigma^2 t^2}{2}}$
Normale multivariée $\mathcal{N}_p(\boldsymbol{m}, \boldsymbol{\Sigma})$	$f(x) = K e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})}$ $K = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}}$ $\boldsymbol{x} = (x_1, ..., x_p)^T \in \mathbb{R}^p$	$\boldsymbol{m}$	$\boldsymbol{\Sigma}$	$e^{i\boldsymbol{u}^T\boldsymbol{m} - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}}$
Khi$_2$ χ^2_ν $\Gamma\left(\frac{1}{2}, \frac{\nu}{2}\right)$	$f(x) = k e^{-\frac{x}{2}} x^{\frac{\nu}{2}-1}$ $k = \dfrac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}$ $\nu \in \mathbb{N}^*,\ x \geq 0$	ν	2ν	$\dfrac{1}{(1-2it)^{\frac{\nu}{2}}}$
Cauchy $c_{\lambda,\alpha}$	$f(x) = \dfrac{1}{\pi\lambda\left(1 + \left(\frac{x-\alpha}{\lambda}\right)^2\right)}$ $\lambda > 0,\ \alpha \in \mathbb{R}$	(-)	(-)	$e^{i\alpha t - \lambda\lvert t \rvert}$
Beta $B(a, b)$	$f(x) = k x^{a-1}(1-x)^{b-1}$ $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ $a > 0,\ b > 0$ $x \in]0, 1[$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	(*)

LOIS DE PROBABILITÉ DISCRÈTES
$\mathbf{m}$: moyenne σ^2 : variance **F. C. :** fonction caractéristique
$p_k = P[X = k]$ $p_{1,...,m} = P[X_1 = k_1, ..., X_m = k_m]$

LOI	**Probabilités**	$\mathbf{m}$	σ^2	**F. C.**
Uniforme	$p_k = \frac{1}{n}$ $k \in \{1, ..., n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\dfrac{e^{it}\left(1 - e^{itn}\right)}{n\left(1 - e^{it}\right)}$
Bernoulli	$p_1 = P[X = 1] = p$ $p_0 = P[X = 0] = q$ $p \in [0, 1] \quad q = 1 - p$	p	pq	$pe^{it} + q$
Binomiale $B(n, p)$	$p_k = C_n^k p^k q^{n-k}$ $p \in [0, 1] \quad q = 1 - p$ $k \in \{0, 1, ..., n\}$	np	npq	$\left(pe^{it} + q\right)^n$
Binomiale négative	$p_k = C_{n+k-1}^{n-1} p^n q^k$ $p \in [0, 1] \quad q = 1 - p$ $k \in \mathbb{N}$	$n\dfrac{q}{p}$	$n\dfrac{q}{p^2}$	$\left(\dfrac{p}{1 - qe^{it}}\right)^n$
Multinomiale	$p_{1,...,m} = \frac{n!}{k_1!...k_m!} p_1^{k_1} ... p_m^{k_m}$ $p_j \in [0, 1] \quad q_j = 1 - p_j$ $k_j \in \{0, 1, \ldots, n\}$ $\sum_{j=1}^m k_j = n \quad \sum_{j=1}^m p_j = 1$	np_j	Variance : $np_j q_j$ Covariance : $-np_j p_k$	$\left(\sum_{j=1}^m p_j e^{it}\right)^n$
Poisson $P(\lambda)$	$p_k = e^{-\lambda} \dfrac{\lambda^k}{k!}$ $\lambda > 0 \quad k \in \mathbb{N}$	λ	λ	$\exp\left[\lambda\left(e^{it} - 1\right)\right]$
Géométrique	$p_k = pq^{k-1}$ $p \in [0, 1] \quad q = 1 - p$ $k \in \mathbb{N}^*$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$	$\dfrac{pe^{it}}{1 - qe^{it}}$