Université Paris I, Panthéon - Sorbonne

MASTER M.A.E.F.

Mise à niveau en Statistique

JEAN-MARC BARDET (UNIVERSITÉ PARIS 1, SAMM)

Plan du cours

1. Rappels de probabilités utiles en statistique inférentielle
   (a) Statistiques descriptives
   (b) Retour sur les variables aléatoires
   (c) Retour sur les vecteurs aléatoires
   (d) Le cas gaussien et les lois associées (Student, Chi2, Fisher)
   (e) Théorèmes limite
2. Estimation paramétrique
   (a) Estimation "‘intuitive"’
   (b) Estimation par maximum de vraisemblance
   (c) Intervalles de confiance
3. Tests paramétriques
   (a) Principe général de test
   (b) Tests "‘intuitifs"’
   (c) Test du rapport de vraisemblance

References

[1] Barbe P. et Ledoux M. (1998) *Probabilité*. EDP Sciences.

[2] Dacunha-Castelle, D. et Duflo, M., *Probabilités et Statistiques (I)*, Masson.

[3] Fourdrinier, D., *Statistique inférentielle*, Dunod.

[4] Milhaud, X., *Statistique*, Belin.

[5] Monfort, A., *Cours de statistique mathématique*, Economica.

[6] Saporta, G., *Probabilités, analyse des données et statistiques*. Technip.

Documents accessibles librement sur internet

References

[1] Site de Philippe Besse à Toulouse: `http://wikistat.fr/`

[2] Site de Ricco Rakotomalala: `http://eric.univ-lyon2.fr/~ricco/cours/index.html`

[3] Cours de Marc Lavielle à l'Ecole Polytechnique: `http://sia.webpopix.org/`

[4] Livre de Davidson-MacKinnon sur le site de R. Davidson: `http://russell.vcharite.univ-mrs.fr/EIE/`

[5] Cours de V. Rivoirad à Dauphine: `https://www.ceremade.dauphine.fr/~rivoirar/Poly-L3-StatMath.pdf`

Feuille n^o 1:

Révisions sur les variables et vecteurs aléatoires

1. **(*)** On considère deux v.a. X et Y définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$. Supposons que X et Y ont la même loi. Est-t-il nécessairement vrai que $\forall \omega \in \Omega, \quad X(\omega) = Y(\omega)$?

2. **(**)** Sur $(\Omega, \mathcal{A}, P)$ un espace de probabilité, on considère une v.a. réelle positive X de fonction de répartition F_X. Déterminer dans les 2 cas suivants l'espérance et la variance de X :
$$\begin{aligned} F_X(t) &= \frac{1}{2}\left(e^t \mathbb{I}_{]-\infty,0[}(t) + (2-e^{-t})\mathbb{I}_{[0,\infty[}(t)\right) ; \\ F_X(t) &= \frac{1}{4}(t+2)\mathbb{I}_{[-1,0[\cup[1,2[}(t) + \frac{3}{4}\mathbb{I}_{[0,1]}(t) + \mathbb{I}_{[2,\infty[}(t). \end{aligned}$$

3. **(*)** On a observé $(X_1, \ldots, X_n)$. Déterminer $\widehat{m} = \underset{m\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (X_i - m)^2$.

4. **(**)** Soit X une v.a. réelle normale centrée réduite et soit Y une v.a. indépendante de X, à valeurs dans $\{-1, 1\}$ telle que $P(Y = 1) = 0.5$. On considère la v.a. $Z = X \cdot Y$. Déterminer la mesure de probabilité de Z, puis celle de $X + Z$. En déduire que la somme de 2 variables gaussiennes peut ne pas être gaussienne.

5. **(**)** Soit X et Y deux v.a. indépendantes définies sur le même espace de probabilité (Ω, P), X suivant une loi de Bernoulli de paramètre p, et Y suivant une loi de Poisson de paramètre λ. On pose pour tout $\omega \in \Omega$,
$$Z(\omega) = \begin{cases} 0 & \text{si} \quad X(\omega) = 0 \\ Y(\omega) & \text{si} \quad X(\omega) = 1 \end{cases}$$
   (a) Montrer que pour tout $\omega \in \Omega \quad Z(\omega) = X(\omega)Y(\omega)$.
   (b) Calculer l'espérance et la variance de Z.

6. **(***)** Soit $(X_1, \ldots, X_n)$ un échantillon de v.a.i.i.d. gaussiennes de moyenne m et de variance $\sigma^2 > 0$.
   (a) On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Déterminer la loi de $\overline{X}_n$.
   (b) Si on note $\mathbb{I} = {}^t(1, 1, \ldots, 1)$, montrer que le projeté orthogonal du vecteur ${}^t(X_1, \ldots, X_n)$ sur le sev de $\mathbb{R}^n$ engendré par $\mathbb{I}$ est $\overline{X}_n\,\mathbb{I}$.
   (c) On note $\overline{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. En utilisant le Théorème de Cochran, démontrer que $(n-1)\,\frac{\overline{\sigma}_n^2}{\sigma^2} \overset{\mathcal{L}}{\sim} \chi^2(n-1)$.

(d) En utilisant le Théorème de Cochran, démontrer que $\overline{X}_n$ et $\overline{\sigma}_n^2$ sont deux v.a. indépendantes. En déduire la loi de $\widehat{T}_n = \dfrac{\sqrt{n}\,(\overline{X}_n - m)}{\sqrt{\overline{\sigma}_n^2}}$.

7. **(**)** On effectue un sondage sur les intentions de vote d'une élection présidentielle. On supposera que chaque électeur à une probabilité p de voter pour le candidat C (p est inconnue). On interroge n personnes choisies au hasard. On note p_n la variable aléatoire qu'est la proportion d'électeurs parmi n désirant voter pour C.

   (a) Donner la loi approchée de p_n.

   (b) Que peut-on dire sur p_n lorsque $n \to +\infty$?

   (c) D'après le sondage, k personnes disent vouloir voter pour C. On note $\hat{p} = k/n$. Déterminer la valeur minimale de n telle que la probabilité de l'événement $\{\mid p_n - p \mid \leq 1\%\}$ soit supérieure à 95%? (on étudiera les cas $\hat{p} \sim 0.5$ et $\hat{p} \sim 0$).

8. **(***)** Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a.i.i.d. à variance finie. Soit $Y_k = \overline{X}_k$ pour $k \geq 1$. Peut-on obtenir la loi faible des grands nombres pour la suite (Y_k) ? La loi forte des grands nombres? Un théorème de limite centrale ? Traiter ensuite le cas particulier où les X_k suivent une loi normale centrée réduite.

Feuille n^o 2:

Estimation et tests paramétriques

1. **(*)** Soit $(X_1, \dots, X_n)$ un échantillon de v.a.i.i.d. Déterminer l'estimateur par maximum de vraisemblance de θ lorsque la loi de X_1 est
   (a) Une loi de Bernoulli de paramètre θ.
   (b) Une loi de Poisson de paramètre θ.
   (c) Une loi normale de paramètre $(\theta, 1)$.
   (d) Une loi exponentielle de paramètre θ.

2. **(**)** Soit $(X_1, \dots, X_n)$ un échantillon de v.a.i.i.d. gaussiennes de moyenne m et de variance $\sigma^2 > 0$. On note $\theta = {}^t(m, \sigma^2)$
   (a) Déterminer $\widehat{\theta} = {}^t(\widehat{m}, \widehat{\sigma}^2)$ l'estimateur de θ par maximum de vraisemblance.
   (b) En utilisant l'exercice 6 précédent, montrer que $\widehat{\theta}$ n'est pas un estimateur sans biais.
   (c) On note $\overline{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$. Avec toujours l'exercice 6 précédent, montrer que le risque quadratique de $\widehat{\sigma}^2$ est inférieur à celui de $\overline{\sigma}_n^2$. Conclusion sur le fait d'être biaisé ou non?

3. **(***)** Soit $(X_1, \dots, X_n)$ un échantillon de v.a.i.i.d. de loi uniforme sur $[0, 2\,\theta]$, où θ est inconnu.
   (a) Déterminer $\mathbb{E}[X_1]$. En déduire que $\widetilde{\theta} = \overline{X}_n$ est un estimateur sans biais et convergent de θ. Déterminer le risque quadratique de $\widetilde{\theta}$.
   (b) Déterminer $\widehat{\theta}$ l'estimateur par maximum de vraisemblance de θ.
   (c) Déterminer la densité de $\widehat{\theta}$. En déduire le risque quadratique de $\widehat{\theta}$. Comparer avec celui de $\widetilde{\theta}$ et conclusion.
   (d) Déterminer un intervalle de confiance à 95% de θ avec $\widetilde{\theta}$ et $\widehat{\theta}$.

4. **(***)** $(X_1, \dots, X_n)$ un échantillon de v.a.i.i.d. de loi dont la densité rapport à la mesure de Lebesgue est:
$$f(x) = \mathbb{I}_{]\theta, \theta+1[}(x).$$
   (a) Montrer que l'estimateur du maximum de vraisemblance pour ce modèle n'est pas unique.
   (b) On pose $\widehat{\theta}_n^{(1)} = \max(X_1, \dots, X_n) - 1$ et $\widehat{\theta}_n^{(2)} = \min(X_1, \dots, X_n)$. Pour chacun de ces estimateurs, déterminer la loi, l'espérance et la variance. Sont-ils indépendants? Convergents?

(c) Déterminer un estimateur de la forme $\widehat{\theta}_n = \alpha\,\widehat{\theta}_n^{(1)} + (1-\alpha)\,\widehat{\theta}_n^{(2)}$, avec $\alpha \in]0,1[$, qui soit sans biais. Cet estimateur a-t-il une variance inférieure à celle des précédents?

5. **(*)** On a lancé 100 fois une pièce de monnaie et obtenu 58 fois pile. Avec un niveau $\alpha = 0.05$, tester si la pièce est équilibrée ? Déterminer la fonction puissance du test. Tester si la pièce est truquée. Aboutit-on au même résultat ?

6. **(*)** Une chaîne des magasins décide d'adopter une nouvelle politique de gestion des stocks pour l'ensemble de ses succursales. Auparavant, le bénéfice mensuel d'une succursale était en moyenne égal à 300000 euros. Après la mise en place de la nouvelle politique pour 100 succursales "pilotes", on observe un bénéfice moyen de 320000 euros avec un écart-type empirique de 20000 euros. Peut-on conclure à l'efficacité de cette politique au niveau 5% ? 1% ?

7. **(**)** Dans le cas d'un échantillon $(X_1, \cdots, X_n)$ de v.a.i.i.d. déterminer le test du rapport de vraisemblance de niveau 5% (statistique de test, région critique), pour les hypothèses:

   (a) $H_0\ : \theta = \theta_0$ contre $H_1\ : \theta = \theta_1 > \theta_0$;

   (b) $H_0\ : \theta = \theta_0$ contre $H_1\ : \theta > \theta_0$;

   (c) $H_0\ : \theta = \theta_0$ contre $H_1\ : \theta \neq \theta_0$;

   (d) $H_0\ : \theta < \theta_0$ contre $H_1\ : \theta > \theta_1 > \theta_0$;

   (e) $H_0\ : \theta \neq \theta_0$ contre $H_1\ : \theta = \theta_0$,

   où $(\theta_0, \theta_1) \in \Theta^2$ sachant que la loi de X_1 est

   (a) une loi de Bernoulli de paramètre $\theta \in]0,1[$;

   (b) une loi uniforme sur $[0, \theta]$ avec $\theta > 0$;

   (c) une loi exponentielle de paramètre $\theta > 0$.