Utilisation en modélisation

Fonction de répartition empirique et ensembles de confiance

1 Problème et son modèle statistique

Dans cet exercice on considère un problème lié à l'utilisation de l'aspirine en médecine. Les données statistiques sont issues de l'article "Heart attack risk found to be cut by taking aspirin" paru dans le New York Times le 27 janvier 1997. Le problème consiste à déterminer l'effet de l'aspirine sur la probabilité d'avoir une crise cardiaque ou une congestion cérébrale. Il y avait deux groupes de patients: un groupe prenant de l'aspirine (11037 patients) et l'autre prenant du placebo (11034 patients). Les médecins qui suivaient ces patients ne savaient pas si le patient prenait de l'aspirine ou du placebo; dans chaque groupe ils ont enregistré les nombres des crises cardiaques et des congestions cérébrales.

1.1 Crise cardiaque

1.1.1 Les données

Le résumé des données concernant les crises cardiaques se trouve dans le tableau suivant

	nombre des crises cardiaques	nombre des patients
groupe prenant de l'aspirine	104	11037
groupe prenant du placebo	189	11034

Tableau 1.

Ce qui est très étonnant dans ces données, c'est un nombre bas de crises cardiaques dans le groupe prenant de l'aspirine. Le rapport

$$\hat{\theta} = \frac{104/11037}{189/11034} = 0.55$$

signifie que l'aspirine divise presque par deux la probabilité de subir une crise cardiaque. Mais la question qui se pose est la suivante: si on répète encore une fois cette étude médicale pourrait-on arriver à la même conclusion?

1.1.2 Le modèle

Pour modéliser les données concernant la crise cardiaque on va utiliser le modèle suivant. À chaque patient, on associe une variable qui prend deux valeurs 0 ou 1. La valeur 0 signifie que le patient n'a pas subi de crise cardiaque et la valeur 1 signifie qu'il en a subi une. Pour le patient i dans le groupe prenant de l'aspirine on notera cette variable par A_i et pour le patient dans le groupe prenant du placebo par P_i. Alors, on dispose de deux échantillons

$$A = (A_1, \ldots, A_m), \quad m = 11037$$

et

$$P = (P_1, \ldots, P_n), \quad n = 11034.$$

Notre hypothèse principale est ce que les v.a. A_i et P_i sont i.i.d. de loi Bernoulli. C'est-à-dire

$$\mathbf{P}(A_i = 1) = p_{as}, \quad \mathbf{P}(A_i = 0) = 1 - p_{as}$$
$$\mathbf{P}(P_i = 1) = p_{pl}, \quad \mathbf{P}(P_i = 0) = 1 - p_{pl}$$

Les paramètres p_{as} et p_{pl} sont inconnus. Ces sont les probabilités qu'un patient a de subir une crise cardiaque dans le groupe prenant de l'aspirine et dans le groupe prenant du placebo.

En se basant sur A est sur P il nous faut construire un estimateur pour le rapport

$$\theta = \frac{p_{as}}{p_{lp}}.$$

Il est facile de voir que l'estimateur du maximum de vraisemblance pour θ est donné par

$$\hat{\theta} = \frac{\sum_{i=1}^{m} A_i/m}{\sum_{i=1}^{n} P_i/n}. \tag{1}$$

Cet estimateur de θ est une v.a. et donc il est entaché d'erreurs. Donc la question qui se pose est la suivante: si on répétait cette expérience encore une fois, dans quel ensemble pourrait se trouver la nouvelle valeur de $\hat{\theta}$ avec une grande probabilité (par exemple 0.95). En statistique, l'idée qui consiste dans l'utilisation d'un ensemble qui couvre le paramètre inconnu avec une grande probabilité est appelée *estimation par ensembles de confiance*. Autrement dit, dans le cas considéré on cherche un intervalle $[\underline{\theta}, \overline{\theta}]$ t.q.

$$\mathbf{P}\big\{\theta \in [\underline{\theta}, \overline{\theta}]\big\} \geq 1 - \alpha \tag{2}$$

où la valeur $1-\alpha$ est appelée *niveau de confiance*. Soulignons que $\underline{\theta}$ et $\overline{\theta}$ sont les fonctions des échantillons A et P. Très souvent, on cherche un intervalle symétrique t.q.

$$\mathbf{P}\big(\theta > \overline{\theta}\big) \leq 1 - \frac{\alpha}{2}, \quad \mathbf{P}\big(\theta < \underline{\theta}\big) \geq \frac{\alpha}{2} \tag{3}$$

Évidement, il existe beaucoup d'intervalles satisfaisant (2) ou (3), et on cherche un intervalle de taille la plus petite possible. Malheureusement, ce problème est très difficile et sa solution n'existe que dans des cas particuliers. En pratique, il y a deux méthodes principales pour construire un intervalle de confiance de taille raisonnablement petite:

- méthode se basant sur le principe du maximum de vraisemblance (tests statistiques)
- méthode qui utilise un estimateur dont on peut calculer ou estimer assez facilement la fonction de répartition.

Dans cet exercice on ne considère que la seconde approche. Nous choisissons l'estimateur $\hat{\theta}$ défini par (1) et supposons un instant que la fonction de répartition de $\hat{\theta} - \theta$

$$F(x, p_{as}, p_{pl}) = \mathbf{P}\big(\hat{\theta} - \theta \leq x\big)$$

est connue. Soulignons que cette fonction de répartition dépend de p_{as}, p_{pl} qui sont inconnus. Si ces paramètres étaient connus, évidemment, il nous suffirait de trouver les quantiles $q_\alpha(p_{as}, p_{pl})$ et $Q_\alpha(p_{as}, p_{pl})$ t.q.

$$F(q_\alpha, p_{as}, p_{pl}) = \frac{\alpha}{2} \quad \text{et} \quad F(Q_\alpha, p_{as}, p_{pl}) = 1 - \frac{\alpha}{2}.$$

pour construire l'intervalle de confiance suivant

$$\underline{\theta} = \hat{\theta} - q_\alpha(p_{as}, p_{pl}), \quad \overline{\theta} = \hat{\theta} + Q_\alpha(p_{as}, p_{pl}).$$

Le problème principal de cette approche consiste à donc trouver la fonction de répartition $F(x, p_{as}, p_{pl})$. Pour calculer cette fonction on va utiliser deux méthodes qui se basent sur

- le théorème de la limite centrale (approximation gaussienne)
- la fonction de répartition empirique (méthode du rééchantillonage)

2 Approximation gaussienne

Le résultat suivant donne la loi asymptotique de $\hat{\theta} - \theta$ lorsque $n, m \to \infty$.

Théorème 1 *Soit*

$$\sigma_{m,n}(p_{as}, p_{pl}) = \frac{p_{as}}{p_{pl}} \sqrt{\frac{1 - p_{as}}{m p_{as}} + \frac{1 - p_{pl}}{n p_{pl}}} \tag{4}$$

alors, lorsque $m \to \infty$ *et* $n \to \infty$

$$\frac{\hat{\theta} - \theta}{\sigma_{m,n}(p_{as}, p_{pl})} \xrightarrow{\mathcal{L}} \xi \tag{5}$$

où ξ *suit une loi gaussienne centrée réduite.*

Démonstration. Soient

$$\hat{p}_{as} = \frac{1}{m} \sum_{i=1}^{m} A_i \qquad \hat{p}_{pl} = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{6}$$

les estimateurs du maximum de vraisemblance des paramètres inconnus p_{as} et p_{pl}. Donc par le théorème de la limite centrale, lorsque $m \to \infty$ et $n \to \infty$

$$\left(\sqrt{m}(\hat{p}_{as} - p_{as}), \sqrt{n}(\hat{p}_{pl} - p_{pl}) \right) \xrightarrow{\mathcal{L}} \left(\xi_1 \sqrt{p_{as}(1 - p_{as})}, \xi_2 \sqrt{p_{pl}(1 - p_{pl})} \right)$$

où ξ_1 et ξ_2 sont les v.a. indépendantes gaussiennes centrées réduites. D'où, à l'aide du développement de Taylor, on obtient (5). □

Ce théorème nous permet de construire un intervalle de confiance de niveau α. Soit t_α le quantile de niveau $\alpha/2$ de loi gaussienne centrée réduite, c'est-à-dire la racine

$$\frac{1}{\sqrt{2\pi}} \int_{t_\alpha}^{\infty} \mathrm{e}^{-u^2/2} \, du = \frac{\alpha}{2} \tag{7}$$

donc on pose

$$\begin{aligned} \underline{\theta} &= \hat{\theta} - \sigma_{n,m}(\hat{p}_{as}, \hat{p}_{pl}) t_\alpha \\ \overline{\theta} &= \hat{\theta} + \sigma_{n,m}(\hat{p}_{as}, \hat{p}_{pl}) t_\alpha. \end{aligned} \tag{8}$$

Théorème 2 *Lorsque* $m \to \infty$ *et* $n \to \infty$

$$\mathbf{P}\big(\theta \notin [\underline{\theta}, \overline{\theta}]\big) \to \alpha$$

Démonstration. Notons, que par la loi des grands nombres

$$\sigma_{n,m}(\hat{p}_{as}, \hat{p}_{pl}) \to \sigma_{m,n}(p_{as}, p_{pl}), \qquad m, n \to \infty.$$

Donc, en utilisant le Théorème 1 et la définition de t_α on achève la démonstration. □

3 Méthode du rééchantillonage

Cette technique récente se base sur l'utilisation intensive des ordinateurs. L'idée principale de cette méthode consiste à *estimer* la fonction de répartition de $\hat{\theta} - \theta$ par la méthode de Monte-Carlo. Dans notre cas, cette méthode se réduit aux étapes suivantes:

- calculer deux fonctions de répartition empiriques

$$F_A(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(A_i \le x), \qquad F_P(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(P_i \le x)$$

pour cela il suffit d'estimer les paramètres inconnus p_{as} et p_{pl} en se basant sur les échantillons A et P

$$\hat{p}_{as} = \frac{1}{m}\sum_{i=1}^{m} A_i \qquad \hat{p}_{pl} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

car

$$F_A(x) = (1-\hat{p}_{as})\mathbf{1}(x \ge 0) + \hat{p}_{as}\mathbf{1}(x \ge 1),$$
$$F_P(x) = (1-\hat{p}_{pl})\mathbf{1}(x \ge 0) + \hat{p}_{as}\mathbf{1}(x \ge 1).$$

- générer N vecteurs indépendants

$$\tilde{A}^l = (\tilde{A}^l_1, \ldots, \tilde{A}^l_m), \qquad \tilde{P}^l = (\tilde{P}^l_1, \ldots, \tilde{P}^l_n) \qquad l = 1, \ldots, N$$

dont les fonctions de répartition sont $F_A(x)$ et $F_P(x)$ respectivement.

- calculer les estimateurs

$$\tilde{\theta}^l = \frac{\sum_{i=1}^{m} \tilde{A}^l_i/m}{\sum_{i=1}^{n} \tilde{P}^l_i/n} \quad l = 1, \ldots, N \tag{9}$$

- calculer la fonction de répartition empirique de $\tilde{\theta}^l - \hat{\theta}$

$$\tilde{F}_N(x) = \frac{1}{N}\sum_{l=1}^{N} \mathbf{1}(\tilde{\theta}^l - \hat{\theta} \le x) \tag{10}$$

et trouver deux quantiles empiriques

$$\tilde{F}_N(Q_\alpha) = 1 - \frac{\alpha}{2}, \qquad \tilde{F}_N(q_\alpha) = \frac{\alpha}{2} \tag{11}$$

- finalement, construire l'intervalle de confiance $[\underline{\theta}, \overline{\theta}]$ avec

$$\underline{\theta} = \hat{\theta} - q_\alpha, \qquad \overline{\theta} = \hat{\theta} + Q_\alpha. \tag{12}$$

3.0.3 Pourquoi le rééchantillonage marche-t-il?

On notera

$$F(x, p_{as}, p_{pl}) = \mathbf{P}\big(\hat{\theta} - \theta \le x\big)$$

Le théorème suivant explique quelle fonction de répartition rend le rééchantillonage.

Théorème 3 *Lorsque* $N \to \infty$

$$\sup_x |\tilde{F}_N(x) - F(x, \hat{p}_{as}, \hat{p}_{pl})| \overset{\text{P.S.}}{\to} 0.$$

Démonstration. Voir le théorème Glivenko-Cantelli. □

Ensuite, par la loi des grands nombres, pour les estimateurs $\hat{p}_{as}$ et $\hat{p}_{pl}$ on a lorsque $m, n \to \infty$

$$\hat{p}_{as} \overset{\text{P.S.}}{\to} p_{as}, \quad \hat{p}_{pl} \overset{\text{P.S.}}{\to} p_{pl}.$$

Il est facile de voir que la fonction de répartition $F(x, p_{as}, p_{pl})$ est continue par rapport à p_{as}, p_{pl}. Donc, lorsque $m, n \to \infty$

$$\sup_x |F(x, \hat{p}_{as}, \hat{p}_{pl}) - F(x, p_{as}, p_{pl})| \overset{\text{P.S.}}{\to} 0.$$

C'est pourquoi

$$\mathbf{P}\big(\theta \notin [\hat{\theta} - q_\alpha, \hat{\theta} + Q_\alpha]\big) \to \alpha, \quad m, n, N \to \infty$$

3.1 Congestion cérébrale

3.1.1 Les données et le modèle

Le résumé de données statistiques concernant de la congestion cérébrale est le suivant

	nombres des congestion cérébrales	nombre des patients
groupe prenant d'aspirine	119	11037
groupe prenant du placebo	98	11034

Tableau 2.

avec

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21.$$

On utilise le même modèle pour modéliser ces données: on suppose qu'il y a deux échantillons

$$A = (A_1, \ldots, A_m), \quad m = 11037$$

et

$$P = (P_1, \ldots, P_n), \quad n = 11034.$$

où les v.a. A_i et P_i sont i.i.d. de loi Bernoulli

$$\mathbf{P}(A_i = 1) = q_{as}, \quad \mathbf{P}(A_i = 1) = 1 - q_{as}$$
$$\mathbf{P}(P_i = 1) = q_{pl}, \quad \mathbf{P}(P_i = 1) = 1 - q_{pl}$$

Les paramètres q_{as} et q_{pl} sont inconnus. Ils sont les probabilités qu'un patient a de subir une congestion cérébrale dans le groupe prenant de l'aspirine et dans le groupe prenant du placebo. Le problème consiste à construire un intervalle de confiance pour le rapport q_{as}/q_{pl} à l'aide des méthodes déjà utilisées pour la crise cardiaque.

4 Simulation avec MATLAB

4.1 Crise cardiaque

4.1.1 Approximation gaussienne

- En se basant sur les données dans le tableau 1, tracer l'approximation gaussienne pour la fonction de répartition de $\hat{\theta} - \theta$, c'est-à-dire la fonction de répartition de

  $$\hat{\theta} + \xi\sigma(\hat{p}_{as}, \hat{p}_{pl})$$

  où ξ suit une loi gaussienne centrée réduite (voir les formules (4) et (6)).

- Indiquer sur le même graphique l'intervalle de confiance pour $\alpha = 0.05$. Utiliser les formules (7) et (8).

4.1.2 Méthode de rééchantillonage

- Estimer la fonction de répartition de $\hat{\theta} - \theta$ par la méthode de rééchantillonage avec $N = 10000$. Utiliser les formules (9)–(10). Tracer cette fonction de répartition.

- Trouver l'intervalle de confiance de niveau $\alpha = 0.05$ (voir (11) et (12)). Indiquer cet intervalle sur le graphique. Représenter vos résultats comme sur la figure 1.

4.2 Congestion cérébrale

1. Faire le même exercice avec les données du tableau 2. Illustrer vos résultats comme sur la figure 2.

2. Discuter la différence entre deux graphiques. Pourquoi l'effet de l'aspirine est bien clair dans le cas de la crise cardiaque? Où la valeur neutre $\theta = 1$ se trouve-t-elle dans chaque cas?

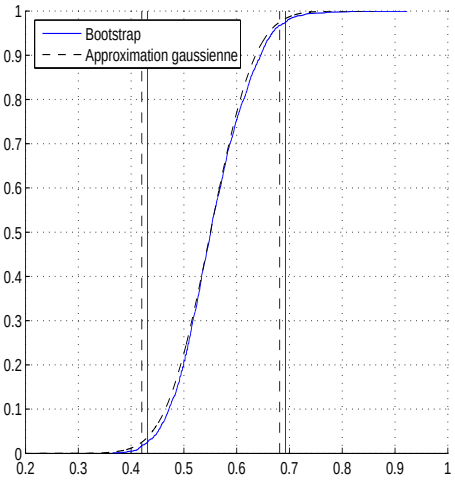

Figure 1: Crise cardiaque.

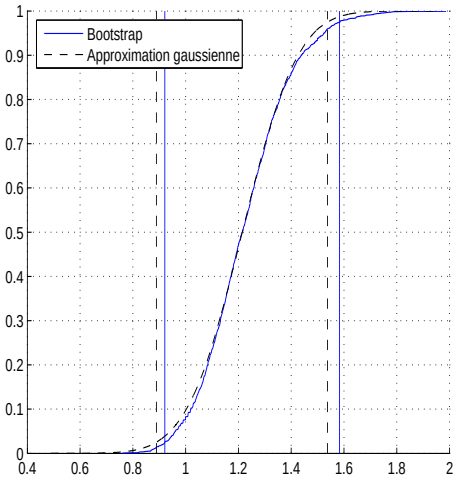

Figure 2: Congestion cérébrale.

Bibliographie

[1] Efron, B. and Tibshirani, R. (1993) *An Introduction to Bootstrap.* Chapmann & Hall.