Ecole Centrale de Nantes

Compte-Rendu TP6 :

Evaluation d’une proportion – Moindres carrés avec contraintes linéaires

Préparé par :

Hugo Salou

Edwin Abou Jaoude

Partie 1. Évaluation d'une proportion

On s'intéresse à un cas concret : une élection entre deux candidats C_1 et C_2. On réalise un sondage auprès d'un grand nombre d'électeurs et on obtient alors une réalisation $\boldsymbol{x} = (x_1, \dots, x_n)$ (sous forme de 0 ou de 1) d'un échantillon $(X_1, \dots, X_n)$ de taille n, où les X_i suivent une loi de Bernoulli $\mathscr{B}(p)$, où p est la proportion des suffrages pour C_1. Dans cette section, on essaie d'estimer la valeur de p, et les incertitudes associées à travers un intervalle de confiance à 95%.

1.1. Construction de l'estimateur

On pose $\bar{p}_k = \frac{1}{k}\sum_{i=1}^k x_i$ une réalisation de la variable $\overline{X}_k = \frac{1}{k}\sum_{i=1}^k X_i$, qui est un estimateur de la proportion p recherchée. Cet estimateur est sans biais, il converge presque sûrement vers p, et converge en L^2.

On sait donc que :

- la dispersion des $\overline{X}_k$ autour de p tend vers 0 pour k grand ;
- la suite $(\bar{p}_k)_{k\in\mathbb{N}}$ converge vers p quelle que soit les valeurs de la réalisation $(x_1, \dots, x_n)$.

1.2. Construction d'un intervalle de confiance

On cherche un intervalle I_k tel que l'on ait la relation $\mathrm{P}(p \in I_k) \simeq 1 - \alpha$. On aimerai un intervalle à 95%, on pose donc $\alpha = 5\%$.

Après calcul, on trouve que :

$$I_k = \overline{X}_k \pm z_{\alpha/2}\sqrt{\frac{\overline{X}_k(1 - \overline{X}_k)}{k}}.$$

Dès lors que les inégalités $k\ \bar{p}_k \geq 5$ et $k\ (1 - \bar{p}_k) \geq 5$ sont vérifiées, on pourra considérer que l'approximation de l'intervalle de confiance $\mathrm{P}(p \in I_k) \simeq 1 - \alpha$ est vérifiée.

1.3. Mise en œuvre sur des échantillons d'intention de vote

Dans cette section, on pose $p = 51\%$, la proportion de candidats à vouloir voter pour le candidat C_1.

On génère une réalisation $\boldsymbol{x}$ d'un échantillon de $N = 10\ 000$ électeurs. Sur la figure ci-après (**Erreur ! Source du renvoi introuvable.**), on donne $(k, \bar{p}_k)$ pour $k = 1\ 000,\ 2\ 000, \dots, 10\ 000$.

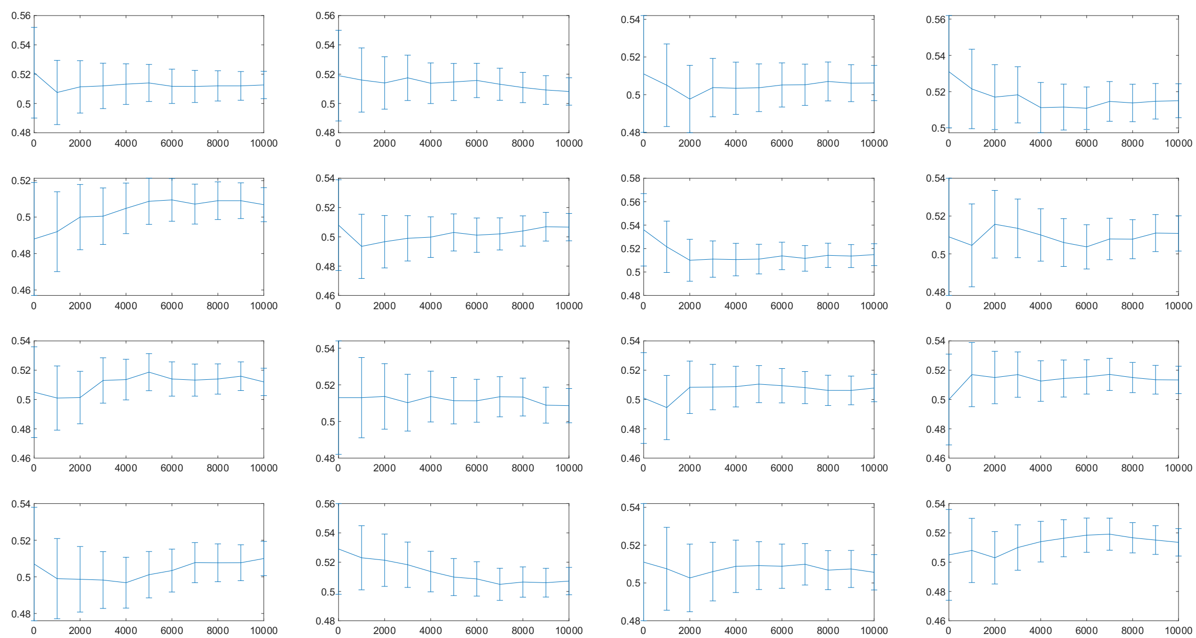

Figure 1. Différentes tentatives d'estimation de la valeur de p

On réalise ce graphe plusieurs fois, avec des réalisations différentes, mais une même valeur de p. On observe que, quand k augmente, nous observons bien les propriétés de l'estimateur :

- ◊ la convergence presque sûre avec la convergence de $\bar{p}_k$ vers p ;
- ◊ la convergence en L^2 avec la diminution de la taille des incertitudes.

Dans la plupart des cas, le candidat C_1 qui a la majorité (51%) des électeurs gagne. Mais, bien plus rarement, il arrive que ce soit le candidat C_2 qui gagne l'élection. On montre un de ces cas dans la **Erreur ! Source du renvoi introuvable.** ci-dessous.

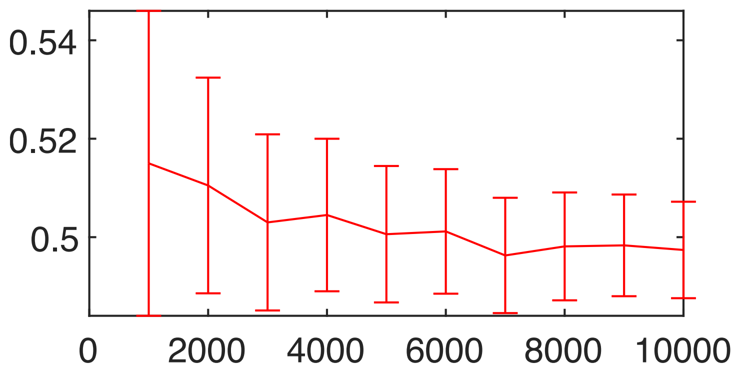

Figure 2. Le candidat C_1 perd l'élection

On considère que C_1 perd l'élection dès lors que $\bar{p}_k < 50\%$. La victoire de C_2, quand elle a lieu, est bien plus souvent pour des plus petites valeurs de k ($k = 1\ 000$, ou $k = 2\ 000$ notamment).

On estime p avec $k = 10\ 000$ réalisations et on vérifie si $p \in I_k$. On réalise ce protocole 100 fois, afin d'estimer le niveau asymptotique de I_k. On obtient des résultats « proches » de 95%, qui varient généralement entre 87% et 100%. On a donc bien un intervalle de confiance de niveau *asymptotique* 95%.

Partie 2. Moindres carrés avec contraintes linéaires

2.1. Moindres Carrés Avec Contraintes Linéaires

Nous allons nous intéresser à un problème de moindres carrés avec des contraintes linéaires uniquement. On souhaite trouver x qui minimise $\frac{1}{2}\|\boldsymbol{Cx}-\boldsymbol{d}\|_2^2$ avec la contrainte $\boldsymbol{A}_{\text{éq}}\boldsymbol{x}=\boldsymbol{b}_{\text{éq}}$, où :

- $\boldsymbol{C} \in \mathcal{M}_{n,m}(\mathbb{R})$ de rang n,
- $\boldsymbol{d}$ est un vecteur.

Le seul point de minimum $\boldsymbol{x}^\star$ sur l'ensemble des contraintes est accompagné d'un multiplicateur de Lagrange $\boldsymbol{\lambda}^\star$ tel que le couple $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ soit solution du système suivant :

$$\begin{pmatrix} \boldsymbol{C}^T\boldsymbol{C} & \boldsymbol{A}_{\text{éq}}^T \\ \boldsymbol{A}_{\text{éq}} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \lambda \end{pmatrix} = \begin{pmatrix} \boldsymbol{C}^T\boldsymbol{d} \\ \boldsymbol{b}_{\text{éq}} \end{pmatrix}$$

La recherche se porte sur le vecteur $\boldsymbol{x}$ qui satisfait au mieux le système $\boldsymbol{Cx}=\boldsymbol{d}$ avec les données suivantes :

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{pmatrix}, \quad \boldsymbol{d} = \begin{pmatrix} 65.5 \\ -142 \\ 82.5 \\ -12 \\ 25 \\ -45 \end{pmatrix}, \quad \boldsymbol{A}_{\text{éq}} = \begin{pmatrix} 4 & 0 & 1 \\ 3 & 1 & 0 \end{pmatrix}, \quad \boldsymbol{b}_{\text{éq}} = \begin{pmatrix} 7 \\ 7 \end{pmatrix}$$

En utilisant la fonction `lsqlineq` dans le script, on obtient la solution $\boldsymbol{x}^\star = (2,1,-1)$, $\boldsymbol{\lambda}^\star = (-1,2)$ et $J(\boldsymbol{x}^\star) = 1.6796\ \times 10^4$ avec le résidu suivant :

$$\text{Résidu} = \begin{pmatrix} -63.5 \\ 144 \\ -82.5 \\ 8 \\ -35 \\ 27 \end{pmatrix}.$$

Il y a une différence entre la solution obtenue avec la fonction de `lsqlin` de Matlab.

```
x =

    1.7476
    1.7573
   -1.1609

>> jval

jval =

   3.3591e+04

>> residual

residual =

  -63.7524
  144.3440
  -81.8814
    8.5715
  -34.7974
   26.5119
```

Figure 3. Résultats en utilisant `sqlin`

2.2. Approximation-interpolation d'un nuage de points

On cherche à déterminer un polynôme P de la forme $P(x) = a_3x^3 + a_2x^2 + a_1x + a_0$ qui interpole les points $(-3,2)$ et $(3,5)$ et qui s'ajuste au mieux, selon la méthode des moindres carres, aux autres points donnés :

x_i	-2	-1	0	1	2
y_i	6	4	5	3	1

L'objectif est de minimiser la fonction $\|\boldsymbol{Cx} - \boldsymbol{d}\|^2$ ou $\boldsymbol{C}$ est une matrice de la forme :

$$\boldsymbol{C} = \begin{bmatrix} 1 & \dots & x_1^3 \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_5^3 \end{bmatrix}$$

Cette matrice est de rang 5 car les x_i sont distincts. Le vecteur $\boldsymbol{x}$ est défini comme :

$$\boldsymbol{x} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

et est de dimension 4. La contrainte est donnée par $\boldsymbol{Ax} = \boldsymbol{b}$ où :

$$\boldsymbol{A} = \begin{bmatrix} 1 & -3 & (-3)^2 & (-3)^3 \\ 1 & 3 & 3^2 & 3^3 \end{bmatrix} \text{ et } \boldsymbol{b} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

Après application de la méthode décrite, le polynôme $P(x)$ trouvé est $P(x) = 0.2622x^3 - 0.0521x^2 - 1.8598x + 3.9691$

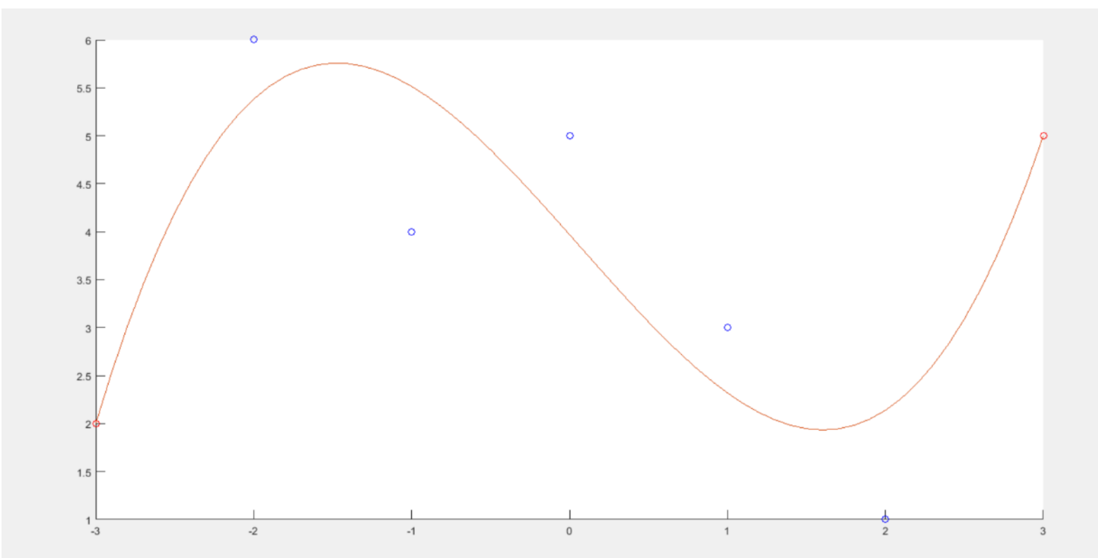

Figure 4. Graphe du polynôme sur l'intervalle $[-3; 3]$