Plaquette 3 – Estimation statistique : fréquences, moyennes et taille d'échantillon

Loi faible des grands nombres — Prépa MP / MPI (2e année)

Exercices

Méthode

Un estimateur est une **moyenne empirique** (ou une fonction continue de moyennes) : la loi faible garantit sa **convergence**, et Bienaymé-Tchebychev donne une **marge d'erreur** valable pour toute loi. Pour dimensionner un échantillon, on résout $\frac{\sigma^2}{n\varepsilon^2} \leq \alpha$, en majorant la variance inconnue ($p(1-p) \leq \frac{1}{4}$, ou une information préalable). On réduit n en réduisant la variance (stratification, information a priori).

Exercice 1 – Taille d'un sondage

Un institut veut estimer la proportion p de personnes favorables à un projet, avec une marge de ± 2 points au risque 5 %.

a) Montrer que $n \geq 12\,500$ personnes suffisent, quelle que soit p.

b) Quelle taille suffirait pour une marge de ± 4 points ?

c) Un concurrent annonce une marge de ± 3 points avec 1 000 personnes. Est-ce justifié par Tchebychev ?

Exercice 2 – Sondage stratifié

Une population comporte 60 % de citadins (proportion favorable $p_A = 0{,}3$) et 40 % de ruraux ($p_B = 0{,}7$).

a) Calculer la proportion globale p et la variance de la fréquence d'un échantillon simple de taille n.

b) On interroge $0{,}6n$ citadins et $0{,}4n$ ruraux, et on estime p par $0{,}6F_A + 0{,}4F_B$. Calculer la variance de cet estimateur.

c) Comparer les tailles nécessaires pour une marge de ± 2 points au risque 5 % (Tchebychev).

Exercice 3 – Estimer l'intensité d'un flux d'appels

Le nombre d'appels reçus par heure suit $P(\lambda)$, indépendamment d'une heure à l'autre. Sur 200 heures, on a reçu 560 appels.

a) Proposer un estimateur de λ et justifier sa convergence.

b) Donner, par Tchebychev, une marge au risque 5 % (on remplacera λ par son estimation dans la variance).

c) Combien d'heures d'observation faudrait-il pour une marge de $\pm 0{,}1$?

Exercice 4 – Estimer un écart type

Une machine produit des pièces dont la longueur X a une moyenne μ et un écart type σ inconnus, avec $\mathbb{E}\left(X^4\right) < +\infty$. On mesure 8 pièces : 98, 102, 101, 99, 100, 103, 97, 100 (en mm).

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

a) Calculer $\overline{x}$ et $s^2 = \frac{1}{7}\sum (x_i - \overline{x})^2$.

b) Justifier que $S_n^2 \to \sigma^2$ en probabilité, puis que $S_n \to \sigma$.

c) L'estimateur S_n est-il sans biais pour σ ?

Exercice 5 – Tirage sans remise dans une population finie

Dans une commune de $N = 2\,000$ électeurs, une proportion p vote pour A. On interroge $n = 400$ électeurs distincts tirés au hasard (sans remise) ; F est la fréquence observée.

a) Rappeler $\mathbb{E}(F)$ et $\mathrm{V}(F)$ (loi hypergéométrique).

b) Donner la marge de Tchebychev au risque 5 % (avec $p(1-p) \leq \frac{1}{4}$), et la comparer à celle d'un tirage avec remise.

c) Que devient la marge si l'on interroge toute la commune ?

Exercice 6 – Fonction de répartition empirique

On observe n temps d'attente indépendants $X_1, \ldots, X_n$ de même loi inconnue, et l'on estime $F(x) = \mathbb{P}(X \leq x)$ par $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$.

a) Montrer que $\hat{F}_n(x) \to F(x)$ en probabilité pour tout x fixé.

b) Combien d'observations garantissent $|\hat{F}_n(x) - F(x)| < 0{,}05$ avec probabilité 0,95 ?

c) Même garantie **simultanément** en 5 valeurs $x_1, \ldots, x_5$: combien d'observations ?

Exercice 7 – Utiliser une information préalable

On veut estimer le taux p de pièces défectueuses à ± 1 point près au risque 5 %. Un audit antérieur garantit $p \leq 0{,}1$.

a) Quelle taille d'échantillon donne Tchebychev sans information ?

b) Montrer que $p \leq 0{,}1$ entraîne $p(1-p) \leq 0{,}09$, et en déduire une nouvelle taille.

c) Quel gain relatif obtient-on ?

Exercice 8 – Comparer deux versions d'un site

Un site teste deux pages d'accueil : n visiteurs voient la version A, n autres la version B, indépendamment. Les taux de clic p_A et p_B sont estimés par F_A et F_B.

a) Montrer que $D_n = F_A - F_B \to p_A - p_B$ en probabilité et majorer $\mathrm{V}(D_n)$.

b) Quel n garantit $|D_n - (p_A - p_B)| < 0{,}02$ avec probabilité 0,95 ?

c) On observe $F_A = 0{,}12$ et $F_B = 0{,}09$ avec cette taille. Peut-on conclure que A est meilleure ?

Exercice 9 – Événements rares : la précision relative

On veut estimer la probabilité $p = 0{,}001$ (ordre de grandeur connu) qu'un composant soit défaillant, avec une erreur **relative** de 10 % au risque 5 %.

a) Traduire l'objectif en une condition $\mathbb{P}(|F_n - p| \geq 0{,}1p) \leq 0{,}05$ et donner n par Tchebychev.

b) Combien de défaillances observera-t-on en moyenne ?

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

c) Pourquoi une marge absolue de ±0,01 serait-elle sans intérêt ici ?

Exercice 10 – Moyenne de lancers de dés

On lance n fois un dé dont on ignore s'il est équilibré ; on sait seulement que ses faces vont de 1 à 6.

a) Montrer que la variance d'une variable à valeurs dans $[1, 6]$ est au plus $\frac{25}{4}$.

b) En déduire n pour estimer l'espérance à ±0,1 au risque 5 %.

c) Si l'on sait le dé équilibré, quelle taille suffirait pour vérifier la moyenne 3,5 avec la même précision ?

Exercice 11 – Que signifie un niveau de 95 % ?

Un institut réalise chaque semaine un sondage indépendant de $n = 12\,500$ personnes et publie l'intervalle $I_k = [F^{(k)} - 0{,}02 ; F^{(k)} + 0{,}02]$ (semaine k), alors que la vraie proportion p reste fixe.

a) Montrer que $\mathbb{P}(p \in I_k) \geq 0{,}95$.

b) Soit C_m la proportion des m premières semaines où l'intervalle contient p. Vers quoi C_m converge-t-elle ?

c) Interpréter le « niveau de confiance ». Un intervalle publié contient-il p avec probabilité 0,95 ?

Exercice 12 – Estimateur par quotient

Un marché vend des fruits au poids. Sur n clients, on relève la dépense D_i et la masse achetée M_i ; les couples (D_i, M_i) sont indépendants et de même loi, avec $\mathbb{E}(M) > 0$ et des variances finies.

a) Montrer que $R_n = \dfrac{\sum D_i}{\sum M_i}$ converge en probabilité vers $\dfrac{\mathbb{E}(D)}{\mathbb{E}(M)}$.

b) Comparer avec $\frac{1}{n}\sum \frac{D_i}{M_i}$ (moyenne des prix au kilo par client) sur l'exemple : client 1, 2 kg pour 6 € ; client 2, 8 kg pour 16 €.

c) Lequel estime le prix moyen d'un kilogramme vendu ?

Exercice 13 – Biais de sélection : demander aux enfants

Dans une ville, les familles ont 1, 2 ou 3 enfants avec probabilités 0,3 ; 0,5 ; 0,2. Pour estimer la taille moyenne des familles, on interroge des **enfants** tirés au hasard parmi tous les enfants de la ville.

a) Calculer la taille moyenne $\mathbb{E}(L)$ d'une famille.

b) Un enfant appartient à une famille de taille ℓ avec probabilité $\dfrac{\ell\,\mathbb{P}(L = \ell)}{\mathbb{E}(L)}$. Vers quoi converge la moyenne des réponses ?

c) Comment corriger l'estimation ?

Exercice 14 – Question sensible et réponse randomisée

Pour estimer la proportion p de fraudeurs, chaque sondé lance en secret une pièce : sur pile, il répond honnêtement à « avez-vous fraudé ? » ; sur face, il relance la pièce

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).

et répond « oui » sur pile, « non » sur face. Soit F_n la fréquence des « oui ».

a) Calculer la probabilité θ d'une réponse « oui » en fonction de p.

b) En déduire un estimateur $\hat{p}_n$ convergent, et majorer sa variance.

c) Combien de sondés pour une marge de $\pm 0{,}05$ au risque 5 % ? Comparer avec une question directe.

Exercice 15 – Synthèse : un sondage au coude-à-coude

Avant un second tour, $n = 1\,000$ électeurs tirés au hasard donnent 52 % des voix au candidat A.

a) Quel théorème justifie que la fréquence observée approche la vraie proportion p ?

b) Donner l'intervalle de Tchebychev au niveau 95 %. Peut-on annoncer la victoire de A ?

c) Quelle taille d'échantillon permettrait, par Tchebychev, une marge de ± 1 point ?

d) Comparer avec l'approximation normale (marge $1{,}96\sqrt{\dfrac{p(1-p)}{n}}$) pour $n = 1\,000$ et pour une marge de ± 1 point.

mathlvl.fr — Tous droits réservés. Toute reproduction, diffusion, revente ou exploitation commerciale, même partielle, est strictement interdite.
Tout contrevenant s'expose à des poursuites en vertu du Code de la propriété intellectuelle (art. L.335-2 et suivants).