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

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

Corrections détaillées

Boîte à outils

Estimer une moyenne. $X_1, \dots, X_n$ indépendantes de même loi, d'espérance μ et de variance σ^2 :

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

— **Taille d'échantillon** pour une marge ε au risque α : $n \geq \dfrac{\sigma^2}{\alpha\varepsilon^2}$.

Estimer une proportion. $\sigma^2 = p(1-p) \leq \frac{1}{4}$, d'où la condition universelle

$$n \geq \frac{1}{4\alpha\varepsilon^2}$$

— Intervalle de Tchebychev au niveau $1-\alpha$: $\left[F_n - \frac{1}{2\sqrt{n\alpha}}, F_n + \frac{1}{2\sqrt{n\alpha}}\right]$.

Estimateurs composés. Si $\overline{X}_n \to \mu$ et $\overline{Y}_n \to \nu$, alors $g(\overline{X}_n, \overline{Y}_n) \to g(\mu, \nu)$ pour g continue (somme, produit, quotient si $\nu \neq 0$, racine).

Variance empirique. $S_n^2 = \frac{1}{n-1}\sum (X_i - \overline{X}_n)^2 \to \sigma^2$ en probabilité (si $\mathbb{E}(X^2)$ est finie et la loi faible s'applique à X^2).

Correction de l'exercice 1 – Taille d'un sondage

a) $nF_n \sim B(n,p)$, $\mathrm{V}(F_n) = \frac{p(1-p)}{n} \leq \frac{1}{4n}$. **Tchebychev** :

$$\mathbb{P}(|F_n - p| \geq 0{,}02) \leq \frac{1}{4n \times 0{,}0004} = \frac{625}{n}$$

— On veut $\frac{625}{n} \leq 0{,}05$, soit $n \geq 12\,500$.

b) Comme $n \propto \frac{1}{\varepsilon^2}$, doubler la marge divise n par 4 (et non par 2) : il suffit de 3 125 personnes.

c) Avec $n = 1\,000$ et $\varepsilon = 0{,}03$: la borne vaut $\frac{1}{4 \times 1\,000 \times 0{,}0009} \approx 0{,}278$. Tchebychev ne garantit qu'un risque de 28 %.

— Le concurrent utilise l'approximation normale (théorème central limite), plus fine mais approchée : $1{,}96 \times \frac{0{,}5}{\sqrt{1\,000}} \approx 0{,}031$.

Réponse. $n \geq 12\,500$; 3 125 pour ±4 points ; non, Tchebychev donne un risque de 0,278.

Correction de l'exercice 2 – Sondage stratifié

a) $p = 0{,}6 \times 0{,}3 + 0{,}4 \times 0{,}7 = 0{,}46$, et $\mathrm{V}(F_n) = \frac{0{,}46 \times 0{,}54}{n} = \frac{0{,}2484}{n}$.

b) F_A et F_B sont indépendantes, de variances $\frac{0{,}21}{0{,}6n}$ et $\frac{0{,}21}{0{,}4n}$:

$$\mathrm{V}(0{,}6F_A + 0{,}4F_B) = 0{,}36 \times \frac{0{,}21}{0{,}6n} + 0{,}16 \times \frac{0{,}21}{0{,}4n} = \frac{0{,}126 + 0{,}084}{n} = \frac{0{,}21}{n}$$

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).

— L'estimateur est sans biais : son espérance vaut $0{,}6p_A + 0{,}4p_B = p$.

c) $n \geq \dfrac{\sigma^2}{\alpha\varepsilon^2}$ avec $\alpha\varepsilon^2 = 0{,}05 \times 0{,}0004 = 2 \times 10^{-5}$:

— simple : $n \geq 12\,420$; stratifié : $n \geq 10\,500$.

— **Lecture** : la stratification supprime la part de variance due aux **différences entre groupes**, ici environ 15 % d'économie.

Réponse. $p = 0{,}46$; variances $\dfrac{0{,}2484}{n}$ et $\dfrac{0{,}21}{n}$; 12 420 contre 10 500.

Correction de l'exercice 3 – Estimer l'intensité d'un flux d'appels

a) $\hat{\lambda} = \overline{X}_n$, moyenne des comptages : **loi faible**, car $\mathbb{E}(X) = \lambda$ et $\mathrm{V}(X) = \lambda$ finie. Ici $\hat{\lambda} = \dfrac{560}{200} = 2{,}8$ appels par heure.

b) $\mathrm{V}(\overline{X}_n) = \dfrac{\lambda}{n}$; on veut $\dfrac{\lambda}{n\varepsilon^2} = 0{,}05$, soit

$$\varepsilon = \sqrt{\frac{\lambda}{0{,}05\,n}} \approx \sqrt{\frac{2{,}8}{10}} \approx 0{,}53$$

— Marge $\pm 0{,}53$ appel par heure : $\lambda \in [2{,}27 ; 3{,}33]$ environ.

c) $n \geq \dfrac{2{,}8}{0{,}05 \times 0{,}01} = 5\,600$ heures (environ 233 jours).

— **Précaution** : remplacer λ par $\hat{\lambda}$ dans la variance est une approximation ; pour une garantie stricte, il faudrait un majorant connu de λ.

Réponse. $\hat{\lambda} = 2{,}8$; marge $\approx \pm 0{,}53$; 5 600 heures.

Correction de l'exercice 4 – Estimer un écart type

a) $\overline{x} = \dfrac{800}{8} = 100$; écarts $-2, 2, 1, -1, 0, 3, -3, 0$, de carrés de somme 28 :

$$s^2 = \frac{28}{7} = 4 \qquad s = 2 \text{ mm}$$

b) On écrit $S_n^2 = \dfrac{n}{n-1}\left(\dfrac{1}{n}\sum X_i^2 - \overline{X}_n^2\right)$.

— **Loi faible** pour X_i et pour X_i^2 (de variance finie grâce à $\mathbb{E}(X^4) < +\infty$) : $\dfrac{1}{n}\sum X_i^2 \to \sigma^2 + \mu^2$ et $\overline{X}_n \to \mu$.

— **Opérations** : $\overline{X}_n^2 \to \mu^2$, $\dfrac{n}{n-1} \to 1$, d'où $S_n^2 \to \sigma^2$; puis $S_n = \sqrt{S_n^2} \to \sigma$ (racine continue).

c) S_n^2 est sans biais, mais **pas** S_n : la racine est concave, et $\mathbb{E}(S_n) \leq \sqrt{\mathbb{E}(S_n^2)} = \sigma$ (Cauchy-Schwarz), avec inégalité stricte dès que S_n n'est pas constante. Le biais disparaît quand $n \to +\infty$.

Réponse. $\overline{x} = 100$, $s = 2$ mm ; $S_n \to \sigma$; S_n biaisé par défaut.

Correction de l'exercice 5 – Tirage sans remise dans une population finie

a) nF suit la loi hypergéométrique : $\mathbb{E}(F) = p$ et

$$\mathrm{V}(F) = \frac{p(1-p)}{n} \times \frac{N-n}{N-1}$$

b) $\mathrm{V}(F) \leq \dfrac{0{,}25}{400} \times \dfrac{1\,600}{1\,999} \approx 5{,}00 \times 10^{-4}$, et la marge ε telle que $\dfrac{\mathrm{V}(F)}{\varepsilon^2} = 0{,}05$ vaut

— sans remise : $\varepsilon \approx \sqrt{\dfrac{5{,}00 \times 10^{-4}}{0{,}05}} \approx 0{,}100$;

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).

— avec remise : $\varepsilon = \sqrt{\frac{0{,}25}{400 \times 0{,}05}} \approx 0{,}112$.

c) $n = N$ annule le facteur $\frac{N-n}{N-1}$: $\mathrm{V}(F) = 0$, la fréquence est exactement p.

— **Lecture** : tant que l'échantillon est petit devant la population, ce facteur est proche de 1 ; c'est pourquoi la précision d'un sondage dépend de n et presque pas de la taille du pays.

Réponse. Facteur $\frac{N-n}{N-1}$; marge $\approx 0{,}100$ contre $0{,}112$; nulle si $n = N$.

Correction de l'exercice 6 – Fonction de répartition empirique

a) Les $1_{\{X_i \le x\}}$ sont indépendantes de loi $B(F(x))$: $\hat{F}_n(x)$ est une fréquence, et la **loi faible** s'applique.

b) **Tchebychev** avec $F(x)(1-F(x)) \le \frac{1}{4}$:

$$\mathbb{P}\left(|\hat{F}_n(x) - F(x)| \ge 0{,}05\right) \le \frac{1}{4n \times 0{,}0025} = \frac{100}{n}$$

— $\frac{100}{n} \le 0{,}05$ donne $n \ge 2\,000$.

c) **Sous-additivité** : la probabilité qu'au moins une des 5 estimations soit mauvaise est au plus $5 \times \frac{100}{n}$. On veut $\frac{500}{n} \le 0{,}05$: $n \ge 10\,000$.

— **Lecture** : contrôler plusieurs estimations à la fois coûte plus cher ; la borne par la somme (dite de Bonferroni) est grossière mais toujours valable.

Réponse. $\hat{F}_n(x) \to F(x)$; 2 000 observations ; 10 000 pour 5 points simultanés.

Correction de l'exercice 7 – Utiliser une information préalable

a) Sans information, $p(1-p) \le \frac{1}{4}$ et

$$n \ge \frac{1/4}{0{,}05 \times 0{,}0001} = 50\,000$$

b) $x \mapsto x(1-x)$ est croissante sur $\left[0, \frac{1}{2}\right]$, donc pour $p \le 0{,}1$: $p(1-p) \le 0{,}1 \times 0{,}9 = 0{,}09$. D'où

— $n \ge \frac{0{,}09}{0{,}05 \times 0{,}0001} = 18\,000$.

c) Le gain est de 64 % : la variance d'une proportion est bien plus faible loin de $\frac{1}{2}$.

— **Piège** : la borne $\frac{1}{4}$ est la pire situation ; pour un taux faible, l'utiliser surdimensionne l'échantillon d'un facteur près de 3.

Réponse. 50 000 sans information ; 18 000 si $p \le 0{,}1$; gain de 64 %.

Correction de l'exercice 8 – Comparer deux versions d'un site

a) **Opérations** : $F_A \to p_A$ et $F_B \to p_B$, donc la différence converge. Les deux fréquences sont **indépendantes** :

$$\mathrm{V}(D_n) = \frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{n} \le \frac{1}{2n}$$

b) **Tchebychev** : $\frac{1}{2n \times 0{,}0004} \le 0{,}05$, soit $n \ge 25\,000$ visiteurs par version.

c) $D_n = 0{,}03$, donc avec probabilité au moins 0,95, $p_A - p_B \in]0{,}01 ; 0{,}05[$: la différence est positive, A est meilleure (au risque 5 %).

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).

— **Remarque** : avec les taux observés (proches de 0,1), la vraie variance est bien inférieure à $\frac{1}{2n}$; en pratique on raffine avec $p(1-p) \approx 0{,}1$.

Réponse. $\mathrm{V}(D_n) \le \frac{1}{2n}$; 25 000 visiteurs par version ; oui, $p_A - p_B > 0{,}01$.

Correction de l'exercice 9 – Événements rares : la précision relative

a) **Tchebychev** avec $\mathrm{V}(F_n) = \frac{p(1-p)}{n}$:

$$\mathbb{P}(|F_n - p| \ge 0{,}1p) \le \frac{p(1-p)}{n \times 0{,}01p^2} = \frac{1-p}{0{,}01\,np}$$

— On veut $\frac{0{,}999}{0{,}01 \times 0{,}001\,n} \le 0{,}05$, soit $n \ge 1\,998\,000$ composants.

b) $np \approx 2\,000$ défaillances : la précision relative dépend surtout du **nombre d'événements observés**, pas du nombre d'essais.

c) Une marge de $\pm 0{,}01$ autour de 0,001 autorise des valeurs dix fois plus grandes que p : elle ne distinguerait même pas p de 0.

— **Lecture** : pour un événement rare, on raisonne en erreur relative, et n croît comme $\frac{1}{p}$.

Réponse. $n \ge 1\,998\,000$; environ 2 000 défaillances observées.

Correction de l'exercice 10 – Moyenne de lancers de dés

a) Pour $X \in [1,6]$, $Y = \frac{X-1}{5} \in [0,1]$ vérifie $\mathrm{V}(Y) \le \frac{1}{4}$ (car $Y^2 \le Y$). Donc

$$\mathrm{V}(X) = 25\,\mathrm{V}(Y) \le \frac{25}{4}$$

— Égalité pour une variable valant 1 ou 6 avec probabilité $\frac{1}{2}$.

b) **Tchebychev** : $\frac{25/4}{n \times 0{,}01} \le 0{,}05$, soit $n \ge 12\,500$ lancers.

c) Avec $\sigma^2 = \frac{35}{12}$: $n \ge \frac{35/12}{0{,}05 \times 0{,}01} \approx 5\,833{,}3$, soit 5 834 lancers.

— **Lecture** : l'ignorance sur la loi se paie : la borne universelle double presque la taille nécessaire.

Réponse. $\mathrm{V}(X) \le \frac{25}{4}$; 12 500 lancers ; 5 834 si le dé est équilibré.

Correction de l'exercice 11 – Que signifie un niveau de 95 % ?

a) $p \notin I_k \Leftrightarrow |F^{(k)} - p| > 0{,}02$, et **Tchebychev** donne

$$\mathbb{P}(p \notin I_k) \le \frac{1}{4 \times 12\,500 \times 0{,}0004} = 0{,}05$$

b) Les indicatrices $1_{\{p \in I_k\}}$ sont indépendantes, de même loi de Bernoulli de paramètre $c = \mathbb{P}(p \in I_1) \ge 0{,}95$. **Loi faible** : $C_m \to c$ en probabilité.

c) Sur un grand nombre de sondages, au moins 95 % des intervalles publiés contiennent p : c'est la méthode qui a un taux de succès garanti.

— Une fois l'intervalle calculé, il contient p ou non (rien d'aléatoire) : la probabilité porte sur la **procédure**, pas sur un intervalle particulier.

Réponse. $\mathbb{P}(p \in I_k) \ge 0{,}95$; $C_m \to c \ge 0{,}95$; le niveau porte sur la procédure.

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).

Correction de l’exercice 12 – Estimateur par quotient

a) $R_n = \dfrac{\overline{D}_n}{\overline{M}_n}$. **Loi faible** : $\overline{D}_n \to \mathbb{E}(D)$ et $\overline{M}_n \to \mathbb{E}(M) > 0$. La fonction $(x, y) \mapsto \frac{x}{y}$ est continue en $(\mathbb{E}(D), \mathbb{E}(M))$:

$$R_n \xrightarrow{\mathbb{P}} \frac{\mathbb{E}(D)}{\mathbb{E}(M)}$$

— (On montre comme pour le produit que $\frac{1}{\overline{M}_n} \to \frac{1}{\mathbb{E}(M)}$, puis on multiplie.)

b) $R_2 = \dfrac{6+16}{2+8} = 2{,}2$ €/kg, alors que la moyenne des prix au kilo vaut $\dfrac{3+2}{2} = 2{,}5$ €/kg.

c) R_n : il pondère chaque client par la masse achetée, c’est le prix moyen d’un kilogramme. La moyenne des quotients estime $\mathbb{E}\left(\frac{D}{M}\right)$, le prix au kilo d’un **client** moyen — une autre quantité.

Réponse. $R_n \to \frac{\mathbb{E}(D)}{\mathbb{E}(M)}$; 2,2 contre 2,5 €/kg ; R_n estime le prix d’un kilogramme vendu.

Correction de l’exercice 13 – Biais de sélection : demander aux enfants

a) $\mathbb{E}(L) = 0{,}3 + 1 + 0{,}6 = 1{,}9$ enfant.

b) **Loi faible** : la moyenne des réponses converge vers l’espérance de la loi **biaisée par la taille** :

$$\sum_{\ell} \ell \, \frac{\ell \, \mathbb{P}(L=\ell)}{\mathbb{E}(L)} = \frac{\mathbb{E}\left(L^2\right)}{\mathbb{E}(L)} = \frac{0{,}3 + 2 + 1{,}8}{1{,}9} = \frac{4{,}1}{1{,}9} \approx 2{,}16$$

— L’enquête surestime la taille moyenne de 14 % : les grandes familles ont plus d’enfants à interroger.

c) On pondère chaque réponse ℓ_i par $\frac{1}{\ell_i}$: la moyenne des $\frac{1}{\ell_i}$ converge vers $\sum_{\ell} \frac{1}{\ell} \cdot \frac{\ell \, \mathbb{P}(L=\ell)}{\mathbb{E}(L)} = \frac{1}{\mathbb{E}(L)}$, d’où l’estimateur convergent $\left(\frac{1}{n}\sum_i \frac{1}{\ell_i}\right)^{-1} \to \mathbb{E}(L)$.

Réponse. $\mathbb{E}(L) = 1{,}9$; les réponses convergent vers $\frac{4{,}1}{1{,}9} \approx 2{,}16$; corriger par la moyenne harmonique.

Correction de l’exercice 14 – Question sensible et réponse randomisée

a) **Probabilités totales** : $\theta = \frac{1}{2}p + \frac{1}{2} \times \frac{1}{2} = \frac{p}{2} + \frac{1}{4}$.

b) $F_n \to \theta$ (loi faible), donc par continuité $\hat{p}_n = 2\left(F_n - \frac{1}{4}\right) \to p$.

$$\mathrm{V}(\hat{p}_n) = 4\,\mathrm{V}(F_n) = \frac{4\theta(1-\theta)}{n} \le \frac{1}{n}$$

c) **Tchebychev** : $\dfrac{1}{n \times 0{,}0025} \le 0{,}05$, soit $n \ge 8\,000$, contre $\dfrac{1}{4 \times 0{,}0025 \times 0{,}05} = 2\,000$ pour une question directe.

— **Compromis** : la confidentialité (une réponse « oui » ne prouve rien) se paie par une variance quatre fois plus grande ; mais une question directe sur un sujet sensible donnerait des réponses biaisées, ce qu’aucun n ne corrige.

Réponse. $\theta = \frac{p}{2} + \frac{1}{4}$; $\hat{p}_n = 2(F_n - \frac{1}{4})$, $\mathrm{V} \le \frac{1}{n}$; 8 000 contre 2 000.

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).

Correction de l’exercice 15 – Synthèse : un sondage au coude-à-coude

a) La **loi faible des grands nombres** : $F_n \to p$ en probabilité, car les réponses sont des variables de Bernoulli indépendantes de paramètre p (tirage avec remise, ou sans remise dans une très grande population).

b) **Tchebychev** : marge ε telle que $\frac{1}{4n\varepsilon^2} = 0{,}05$:

$$\varepsilon = \frac{1}{2\sqrt{0{,}05 \times 1\,000}} = \frac{1}{2\sqrt{50}} \approx 0{,}071$$

— Intervalle $[0{,}449 ; 0{,}591]$: il contient des valeurs inférieures à $\frac{1}{2}$, on ne peut pas annoncer la victoire.

c) $n \geq \frac{1}{4 \times 0{,}05 \times 0{,}0001} = 50\,000$.

d) **Approximation normale** : avec $p(1-p) \leq \frac{1}{4}$, la marge vaut $1{,}96 \times \frac{0{,}5}{\sqrt{1\,000}} \approx 0{,}031$, intervalle $[0{,}489 ; 0{,}551]$ — toujours pas concluant. Pour ± 1 point :
$n \geq \left(\frac{1{,}96 \times 0{,}5}{0{,}01}\right)^2 = 9\,604$.

— **Bilan** : Tchebychev est garanti mais pessimiste (facteur 5 sur la taille) ; à 52 %, il faut de toute façon plusieurs milliers de personnes pour trancher.

Réponse. Marge $\approx \pm 0{,}071$ (non concluant) ; 50 000 par Tchebychev ; normale : $\pm 0{,}031$ et 9 604.

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).