STA240 : Estimation parametrique

1 Estimation ponctuelle

- Pour un paramètre inconnu, un estimateur est une fonction des données, qui prend des valeurs proches de ce paramètre. Il est *sans biais* si son espérance est égale au paramètre. Il est *convergent* si la probabilité qu'il prenne des valeurs à distance au plus ε du paramètre, tend vers 1 quand la taille de l'échantillon tend vers l'infini.
- La *fréquence empirique* d'un événement est un estimateur sans biais et convergent de la probabilité de cet événement.
- La *moyenne empirique* d'un échantillon est un estimateur sans biais et convergent de l'espérance théorique des variables.
- La *variance empirique* d'un échantillon est un estimateur convergent de la variance théorique des variables. On obtient un estimateur sans biais en multipliant la variance empirique par $n/(n-1)$, où n est la taille de l'échantillon.

Exercice 1. On considère l'échantillon statistique $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. Calculer sa moyenne et sa variance empiriques.

   On trouve :
   $$\overline{x} = \frac{6}{9} = \frac{2}{3} \quad et \quad s_x^2 = \frac{4}{9} .$$

2. En supposant que les données de cet échantillon sont des réalisations d'une variable de loi inconnue, donner une estimation non biaisée de l'espérance et de la variance de cette loi.

   La moyenne empirique ($2/3$*) est une estimation non biaisée de l'espérance. On obtient une estimation non biaisée de la variance en multipliant* s_x^2 *par* $9/8$ *: on trouve* $1/2$*.*

3. On choisit de modéliser les valeurs de cet échantillon par une loi binomiale $\mathcal{B}(2, p)$. Utiliser la moyenne empirique pour proposer une estimation ponctuelle pour p.

   L'espérance de la loi $\mathcal{B}(2, p)$ *est* $2p$*. Elle est estimée par la moyenne empirique (ici :* $2/3$*). Donc la probabilité* p *peut être estimée par :*
   $$\frac{2/3}{2} = \frac{1}{3} .$$

4. Avec le même modèle, utiliser la variance empirique pour proposer une autre estimation de p.

   La variance de la loi $\mathcal{B}(2, p)$ *est* $2p(1-p)$*. Elle est estimée par* $1/2$*. On obtient une estimation de* p *en résolvant l'équation* $2p(1-p) = 1/2$*, dont la solution est* $p = 1/2$*.*

5. On choisit de modéliser les valeurs de cet échantillon par une loi de Poisson $\mathcal{P}(\lambda)$, qui a pour espérance λ. Quelle estimation ponctuelle proposez-vous pour λ ?

   On estime λ par la moyenne empirique, $2/3$.

Exercice 2. On considère l'échantillon statistique

$$(1, 3, 2, 3, 2, 2, 0, 2, 3, 1) \ .$$

1. En supposant que les variables de cet échantillon sont des réalisations d'une variable de loi inconnue, donner une estimation non biaisée de l'espérance et de la variance de cette loi.
2. On choisit de modéliser les valeurs de cet échantillon par une loi binomiale $\mathcal{B}(3, p)$. Utiliser la moyenne empirique pour proposer une estimation ponctuelle pour p.

Exercice 3. On considère l'échantillon statistique

$$(1.2, 0.2, 1.6, 1.1, 0.9, 0.3, 0.7, 0.1, 0.4) \ .$$

1. On choisit de modéliser les valeurs de cet échantillon par une loi uniforme sur l'intervalle $[0, \theta]$. Quelle estimation ponctuelle proposez-vous pour θ ?
2. On choisit de modéliser les valeurs de cet échantillon par une loi normale $\mathcal{N}(\mu, \sigma^2)$. Quelle estimation ponctuelle proposez-vous pour μ et σ^2 ?

2 Intervalles de confiance pour un échantillon gaussien

Un échantillon gaussien est un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$. On note :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2 \ ,$$

la moyenne et la variance empiriques de l'échantillon.

- Si la variance théorique σ^2 est *connue*, on obtient un intervalle de confiance de niveau $1-\alpha$ pour μ par :

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}} \ ; \ \overline{X} + u_\alpha \frac{\sqrt{\sigma^2}}{\sqrt{n}}\right] \ ,$$

  où u_α est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0, 1)$.

- Si la variance théorique σ^2 est *inconnue*, on obtient un intervalle de confiance de niveau $1-\alpha$ pour μ par :

$$\left[\overline{X} - t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}} \; ; \; \overline{X} + t_\alpha \frac{\sqrt{S^2}}{\sqrt{n-1}}\right] ,$$

où t_α est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre $n-1$.

- Si la variance théorique σ^2 est *inconnue*, on obtient un intervalle de confiance de niveau $1-\alpha$ pour σ^2 par :

$$\left[\frac{nS^2}{v_\alpha} \; ; \; \frac{nS^2}{u_\alpha}\right] ,$$

où u_α est le quantile d'ordre $\alpha/2$ de la loi de khi-deux de paramètre $n-1$, et v_α est son quantile d'ordre $1-\alpha/2$.

Exercice 4. La force de compression d'un type de béton est modélisée par une variable gaussienne d'espérance μ et de variance σ^2. L'unité de mesure est le *psi* (pound per square inch). Dans les questions de 1. à 4., on supposera la variance σ^2 connue et égale à 1000. Sur un échantillon de 12 mesures, on a observé une moyenne empirique de 3250 psi.

1. Donner un intervalle de confiance de niveau 0.95 pour μ.

   Ici, $\alpha = 0.05$ et $1-\alpha/2 = 0.975$. Le quantile d'ordre 0.975 de la loi $\mathcal{N}(0,1)$ est 1.96. L'intervalle de confiance est :

$$\left[3250 - 1.96\frac{\sqrt{1000}}{\sqrt{12}} \; ; \; 3250 + 1.96\frac{\sqrt{1000}}{\sqrt{12}}\right] \; = \; [3232\,;\,3268] .$$

   Il est inutile de donner plus de chiffres que n'en a la moyenne empirique. On arrondit la borne inférieure par défaut, la borne supérieure par excès ; ainsi l'arrondi ne peut qu'agrandir l'intervalle, et on est assuré que le niveau de confiance de l'intervalle donné est au moins égal à 0.95.

2. Donner un intervalle de confiance de niveau 0.99 pour μ. Comparer sa largeur avec celle de l'intervalle précédent.

   Ici, $\alpha = 0.01$ et $1-\alpha/2 = 0.995$.Le quantile d'ordre 0.995 de la loi $\mathcal{N}(0,1)$ est 2.5758. L'intervalle de confiance est :

$$\left[3250 - 2.5758\frac{\sqrt{1000}}{\sqrt{12}} \; ; \; 3250 + 2.5758\frac{\sqrt{1000}}{\sqrt{12}}\right] \; = \; [3226\,;\,3274] .$$

   L'intervalle est plus large que le précédent. Plus la probabilité que la moyenne appartienne à l'intervalle est grande (0.99 au lieu de 0.95), plus cet intervalle doit être large. Si on veut avoir plus confiance dans l'intervalle, il faut accepter qu'il soit moins précis.

3. Si avec le même échantillon on donnait un intervalle de confiance de largeur 30 psi, quel serait son niveau de confiance ?

   La largeur de l'intervalle de confiance de niveau $1-\alpha$ *est :*

$$2u_\alpha \frac{\sqrt{1000}}{\sqrt{12}} \ .$$

   Si cette largeur est égale à 30*, on obtient :*

$$u_\alpha = \frac{30\sqrt{12}}{2\sqrt{1000}} = 1.6432 \ .$$

   Cette valeur est le quantile d'ordre $0.9498 = 1-\alpha/2$ *de la loi* $\mathcal{N}(0,1)$*. Donc* $\alpha = 0.1003$ *et* $1-\alpha = 0.8997$*.*

4. On souhaite maintenant estimer μ avec une précision de ± 15 psi, avec un niveau de confiance de 0.95. Quelle taille minimum doit avoir l'échantillon ?

   Pour un échantillon de taille n*, La précision de l'intervalle de confiance de niveau* 0.95 *est :*

$$\pm 1.96 \frac{\sqrt{1000}}{\sqrt{n}} \ .$$

   Si elle est égale à 15*, on obtient :*

$$n = \left(\frac{1.96\sqrt{1000}}{15} \right)^2 = 17.07 \ .$$

   L'échantillon doit donc être de taille 18 *au moins.*

5. La variance théorique est désormais supposée inconnue. On dispose de la donnée suivante (sur le même échantillon de taille 12) :

$$\sum_{i=1}^{12} x_i^2 = 126761700 \ .$$

   Donnez pour μ un intervalle de confiance de niveau 0.95 et comparez-le avec celui de la question 1, puis un intervalle de confiance de niveau 0.99 et comparez-le avec celui de la question 2.

   La variance estimée est :

$$s^2 = \frac{1}{12} \times 126761700 - (3250)^2 = 975 \ .$$

   Le quantile d'ordre 0.975 *de la loi de Student* $\mathcal{T}(n–1)$ *est* 2.201*, le quantile d'ordre* 0.995 *est* 3.106*. L'intervalle de confiance de niveau* 0.95 *est :*

$$\left[3250 - 2.201 \frac{\sqrt{975}}{\sqrt{11}} \ ; \ 3250 + 2.201 \frac{\sqrt{975}}{\sqrt{11}} \right] \ = \ [3229 \, ; \, 3271] \ .$$

L'intervalle de confiance de niveau 0.99 *est :*

$$\left[3250 - 3.106\frac{\sqrt{975}}{\sqrt{11}}\ ;\ 3250 + 3.106\frac{\sqrt{975}}{\sqrt{11}}\right] \ = \ [3220\,;\,3280]\ .$$

À niveau de confiance égal, et bien que la variance estimée soit inférieure à la variance théorique, l'intervalle de confiance calculé avec la loi de Student (variance supposée inconnue) est plus large, donc moins précis, que celui calculé avec la loi normale (variance connue). Cela tient au fait que les lois de Student sont plus dispersées que la loi normale $\mathcal{N}(0,1)$ *: l'intervalle contenant* 95% *des valeurs pour la loi* $\mathcal{T}(11)$ *est* $[-2.201\ ;\ +2.201]$*, au lieu de* $[-1.96\ ;\ +1.96]$ *pour la loi* $\mathcal{N}(0,1)$*. Il est raisonnable de s'attendre à une moins grande précision quand on dispose de moins d'information sur le modèle.*

6. Donner un intervalle de confiance de niveau 0.95 pour la variance, et pour l'écart-type.

   Le quantile d'ordre 0.025 *pour la loi de khi-deux* $\mathcal{X}^2(11)$ *est* $u_\alpha = 3.816$*. Le quantile d'ordre* 0.975 *est* $v_\alpha = 21.92$*. L'intervalle de confiance de niveau* 0.95 *pour la variance est :*

   $$\left[\frac{12 \times 975}{21.92}\ ;\ \frac{12 \times 975}{3.816}\right] \ = \ [533\,;\,3067]\ .$$

   En prenant la racine carrée des deux bornes, on obtient un intervalle de confiance pour l'écart-type :

   $$\left[\sqrt{\frac{12 \times 975}{21.92}}\ ;\ \sqrt{\frac{12 \times 975}{3.816}}\right] \ = \ [23.1\,;\,55.4]\ .$$

   Les intervalles de confiance pour la variance ou l'écart-type pour de petits échantillons sont en général très imprécis.

Exercice 5. On a mesuré le poids de raisin produit par pied sur 10 pieds pris au hasard dans une vigne. On a obtenu les résultats suivants exprimés en kilogrammes :

$$2.4 \quad 3.4 \quad 3.6 \quad 4.1 \quad 4.3 \quad 4.7 \quad 5.4 \quad 5.9 \quad 6.5 \quad 6.9\ .$$

On modélise le poids de raisin produit par une souche de cette vigne par une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$.

1. Calculer la moyenne et la variance empiriques de l'échantillon.
2. Donner un intervalle de confiance de niveau 0.95 pour μ.
3. Donner un intervalle de confiance de niveau 0.95 pour σ^2.
4. On suppose désormais que l'écart-type des productions par pied est connu et égal à 1.4. Donner un intervalle de confiance de niveau 0.95 pour μ.

5. Quel nombre de pieds au minimum devrait-on observer pour estimer μ au niveau de confiance 0.99 avec une précision de plus ou moins 500 grammes ?

Exercice 6. Une étude faite sur la vitesse coronarienne a donné les résultats suivants sur 18 individus :

$$75,\ 77,\ 78,\ 77,\ 77,\ 72,\ 72,\ 72,\ 70,\ 71,\ 69,\ 69,\ 68,\ 66,\ 64,\ 66,\ 62,\ 61.$$

On modélise les valeurs de cet échantillon par une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma^2)$, où μ et σ^2 sont deux paramètres a priori inconnus.

1. Calculer la moyenne et la variance de l'échantillon.
2. Calculer les intervalles de confiance de μ aux niveaux 0.95, 0.98 et 0.99.
3. Calculer les intervalles de confiance de σ^2 aux niveaux 0.95, 0.98 et 0.99.
4. Que seraient les intervalles de confiance de μ, si on supposait que la variance σ^2 était connue et égale à 26 ?

Exercice 7. Un laboratoire utilise un appareil de mesure optique destiné à mesurer la concentration des solutions de fluoresceïne. Les résultats des mesures sont modélisés par une variable aléatoire normale dont l'espérance est égale à la concentration réelle de la solution, et l'écart-type, garanti par le constructeur, est connu : $\sigma = 0.05$.

1. On effectue 9 mesures à partir d'une solution donnée. La moyenne empirique des 9 mesures est 4.38 mg/l. Donner un intervalle de confiance pour la concentration réelle de la solution, au niveau de confiance 0.99.
2. Pour le même échantillon, quel est le niveau de confiance de l'intervalle $[4.36\ ;\ 4.40]$?
3. Quelle devrait être la taille de l'échantillon pour connaître la concentration réelle de la solution, au niveau de confiance 0.99, avec une précision de ± 0.01 mg/l ?
4. Sur le même échantillon de 9 mesures, on a observé un écart-type empirique de 0.08 mg/l. Donner un intervalle de confiance pour l'écart-type réel, de niveau de confiance 0.99. Que pensez-vous de la garantie du constructeur ?
5. Reprendre la première question, en supposant cette fois que l'écart-type de la loi des mesures est inconnu, et estimé par l'écart-type empirique.

Exercice 8. On désire estimer la production d'une nouvelle espèce de pommier. On modélise la production d'un pommier de cette espèce par une loi normale d'espérance μ et d'écart-type σ inconnus.

1. Sur un échantillon de 15 pommiers, on a observé une récolte moyenne de 52 kg avec un écart-type de 5 kg. Donner un intervalle de confiance pour la production moyenne des pommiers de cette espèce, de niveau 0.95, puis 0.99.
2. Donner un intervalle de confiance pour l'écart-type σ, de niveau 0.95.

3 Int. de conf. d'une espérance pour un grand échantillon

Pour un grand échantillon, on obtient un intervalle de confiance de niveau approché $1-\alpha$ pour l'espérance par :

$$\left[\overline{X} - u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}} \; ; \; \overline{X} + u_\alpha \frac{\sqrt{S^2}}{\sqrt{n}}\right] ,$$

où u_α est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Exercice 9. On a effectué 90 mesures de concentration d'une solution de fluoresceïne. On a observé une moyenne empirique de 4.38 mg/l et un écart-type empirique de 0.08 mg/l. Donner un intervalle de confiance pour la concentration réelle de la solution, au niveaux de confiance 0.95 et 0.99.

Le quantile d'ordre 0.975 *de la loi* $\mathcal{N}(0,1)$ *est* 1.96. *L'intervalle de confiance de niveau* 0.95 *est :*

$$\left[4.38 - 1.96\frac{0.08}{\sqrt{90}} \; ; \; 4.38 + 1.96\frac{0.08}{\sqrt{90}}\right] \; = \; [4.363\,;\, 4.397] \; .$$

Le quantile d'ordre 0.995 *de la loi* $\mathcal{N}(0,1)$ *est* 2.5758. *L'intervalle de confiance de niveau* 0.99 *est :*

$$\left[4.38 - 2.5758\frac{0.08}{\sqrt{90}} \; ; \; 4.38 + 2.5758\frac{0.08}{\sqrt{90}}\right] \; = \; [4.358\,;\, 4.402] \; .$$

Exercice 10. On désire estimer la production d'une nouvelle espèce de pommier. Sur un échantillon de 80 pommiers, on observe une récolte moyenne de 51.5 kg, avec un écart-type de 4.5 kg. Donner un intervalle de confiance pour la production moyenne des pommiers de cette espèce, de niveau 0.95, puis 0.99.

Exercice 11. On a mesuré la longueur en millimètres de 152 œufs de coucou, et obtenu une moyenne empirique de 40.8 mm, pour une variance empirique de 14.7 mm^2. Donner un intervalle de confiance pour la longueur moyenne d'un œuf de coucou, au niveau de confiance 0.95, puis 0.98, puis 0.99.

Exercice 12. On a mesuré la longueur de 150 coquilles de noix et obtenu une moyenne empirique de 27.6 mm, pour un écart-type empirique de 3.7 mm. Donner un intervalle de confiance pour la longueur moyenne d'une coquille de noix, au niveau de confiance 0.99, puis 0.999.

Exercice 13. On administre des somnifères à deux groupes de malades A et B comprenant 50 et 100 individus. Le groupe A reçoit un nouveau somnifère, le groupe B reçoit l'ancien. Les patients du groupe A ont dormi 7.82 heures en moyenne avec un écart-type de 0.24 h ; ceux du groupe B ont dormi 6.75 heures en moyenne avec un écart-type de 0.30 h.

1. Calculer l'intervalle de confiance pour le nombre moyen d'heures de sommeil d'un patient recevant le nouveau somnifère, aux niveaux 0.90, puis 0.95 et 0.99.
2. Même question pour un patient recevant l'ancien somnifère.
3. Pensez-vous que le nouveau somnifère soit plus efficace que l'ancien ?

4 Int. de conf. d'une probabilité pour un grand échantillon

Pour un grand échantillon binaire, on obtient un intervalle de confiance de niveau approché $1-\alpha$ pour la probabilité de l'événement par :

$$\left[\overline{X} - u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}} \; ; \; \overline{X} + u_\alpha \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}}\right] ,$$

où n est la taille de l'échantillon, $\overline{X}$ est la fréquence empirique de l'événement et u_α est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Exercice 14. Afin d'étudier l'influence des rayons X sur la spermatogénèse de Bombyx Mori, on a irradié des mâles au deuxième jour et au quatrième jour du stade larvaire ; ces mâles ont été accouplés avec des femelles non irradiées. On a compté le nombre d'œufs fertiles dans la ponte des femelles, et on a obtenu 4998 œufs fertiles pour 5646 œufs récoltés en tout. On a aussi accouplé des mâles et des femelles non irradiés, avec un résultat de 5834 œufs fertiles sur 6221 œufs récoltés.

1. Donner un intervalle de confiance de niveau 0.95 pour la proportion d'œufs fertiles après irradiation des mâles.

   La fréquence empirique des œufs fertiles après irradiation des mâles est :

   $$F = \frac{4998}{5646} = 0.885 .$$

   L'intervalle de confiance de niveau 0.95 *est :*

   $$\left[0.885 - 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}} \; ; \; 0.885 + 1.96\frac{\sqrt{0.885(1-0.885)}}{\sqrt{5646}}\right]$$

   $$= \; [0.876\,;\, 0.894] .$$

2. Donner un intervalle de confiance de niveau 0.95 pour la proportion d'œufs fertiles de couples non irradiés.

   La fréquence empirique des œufs fertiles parmi les couples non irradiés est :

   $$F = \frac{5834}{6221} = 0.938 .$$

L'intervalle de confiance de niveau 0.95 *est :*

$$\left[0.938 - 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\ ;\ 0.938 + 1.96\frac{\sqrt{0.938(1-0.938)}}{\sqrt{6221}}\right]$$

$$= \quad [0.931\ ;\ 0.944]\ .$$

3. Que pensez-vous de l'influence de l'irradiation sur la fertilité des œufs ?

   Les deux intervalles de confiance ont une intersection vide ; la proportion d'œufs fertiles est donc significativement plus basse pour les mâles irradiés.

Exercice 15. On a observé un échantillon de taille $n = 500$ d'adolescents de 15 ans, dans lequel 210 présentent un surpoids. Soit p la proportion d'adolescents de 15 ans qui présentent un surpoids. Donner un intervalle de confiance pour p, aux niveaux de confiance 0.95 et 0.99.

Exercice 16. Une clinique a proposé une nouvelle opération chirurgicale, et a connu 40 échecs, sur 200 tentatives. On note p le pourcentage de réussite de cette nouvelle opération.

1. Quelle estimation de p proposez-vous ?
2. En utilisant l'approximation normale, donner un intervalle de confiance pour p de niveau de confiance 0.95.
3. Combien d'opérations la clinique devrait-elle réaliser pour connaître le pourcentage de réussite avec une précision de plus ou moins 1%, au niveau de confiance 0.95 ?

Exercice 17. Soient $X1, \ldots, Xn$ une suite de variables aléatoires indépendantes et identiquement distribuées de loi de Bernoulli $p \in [0, 1]$. On pose $\hat{p}_n = 1/n \sum X_i$.

1. Montrer l'inégalité $Var(\hat{p}_n) \leq 1/4n$
2. Un institut de sondage souhaite estimer avec une précision de 3 points (à droite et à gauche) la probabilité qu'un individu vote pour le maire actuel aux prochaines elections. Combien de personnes est-il nécessaire de sonder ?
3. Sur un échantillon représentatif de 1000 personnes, on rapporte les avis favorables pour un homme politique. En novembre, il y avait 38% d'avis favorables, en décembre 36%. Un editorialiste dans son journal prend très au sérieux cette chute de 2 points d'un futur candidat à la présidentielle ! Confirmer ou infirmer la position du journaliste.