



Notes on Probability and Statistics

Yehor KOROTENKO

December 5, 2024

Propriétés de la Moyenne Empirique (LGN)

La moyenne empirique tend en probabilité vers l'espérance quand le nombre de répétitions indépendantes tend vers l'infini.

Si $(X_n)_{n \geq 1}$ est une suite de v.a. indépendantes et de même loi et de carré intégrable,

$$\text{alors } \lim_{n \rightarrow \infty} P \left(\left| \frac{X_1 + X_2 + \dots + X_n}{n} - \mathbb{E}[X_1] \right| > \varepsilon \right) = 0.$$

$$\text{On encore: } \forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} P \left(\left| \frac{X_1 + \dots + X_n}{n} - \mathbb{E}[X_1] \right| \leq \varepsilon \right) = 1$$

Preuve:

1. $\mathbb{E} \left[\frac{X_1 + \dots + X_n}{n} \right] = \mathbb{E}[X_1]$
2. $\text{Var} \left(\frac{X_1 + \dots + X_n}{n} \right) = \frac{\text{Var}(X_1)}{n}$
3. Inégalité de Bienaymé-Tchebychev.

Exemples d'utilisation de la LGN pour estimer des paramètres

1. Poids de saumons en fonction d'une alimentation

Si on prélève au hasard X : le poids du saumon $\in \{q_1, \dots, q_n\}$,
 $\forall j \in \{1, \dots, m\}, P(X = q_j) = \frac{\# \text{saumon de poids } q_j}{\# \text{total des saumons}} = p(q_j)$ germe de la loi de X (inconnu)
 $E(X) = \sum_{j=1}^m q_j \cdot p(q_j)$ (inconnu) Estimation du poids moyen $\theta = \mathbb{E}[X]$
 n prélèvements au hasard $\rightarrow X_1, \dots, X_n$
 $\hat{\theta} = \frac{X_1 + \dots + X_n}{n}$ estimation sans biais de θ et consistante.

2. Probabilité d'observer au moins un oiseau d'une espèce rare

dans le territoire dans une journée. A l'aide de données disant, sur les jours $1, \dots, n$ si l'espèce a été observé ou pas.

Pour le jour i , $X_i \sim \mathcal{B}(p)$ $X_i = 0$ si on observe pas l'espèce p inconnu, et 1 sinon.

$\hat{p} = \frac{X_1 + \dots + X_n}{n}$: proportion des jours parmi les n où on a observé l'espèce.

$p = \mathbb{E}[X_1]$

Modele suite: Si on peut supposer que les X_i sont indépendantes et de même loi, alors $\hat{p}$ est sans biais et consistant. Généralité: si $p = P(A) = \mathbb{E}[1_A]$

3. Même contexte, mais on a des données de comptage

Y_i : nombre d'oiseaux observés de l'espèce.

On veut estimer $P(Y_i = 0)$. On peut faire comme précédemment, en définissant :

$$X_i = 1_{Y_i=0}$$

Autre méthode:

- Modèle : $Y_1, \dots, Y_n$ sont des variables aléatoires indépendantes, de loi $P(\lambda)$, $\lambda > 0$.
- $P(Y_i = 0) = e^{-\lambda}$
- $\hat{\lambda} = \frac{Y_1 + \dots + Y_n}{n}$ (estimateur sans biais et consistant de λ).
- $\tilde{p} = e^{-\hat{\lambda}} = e^{-\frac{Y_1 + \dots + Y_n}{n}}$ est un estimateur consistant de $P(Y_i = 0)$.
- Les estimateurs sont utilisés dans les procédures statistiques dont **tests statistiques**.

Test Statistique : Exemples

1. La pièce est-elle sans biais ?

Expérience : 1000 lancers.

i -ème lancer de pièce :

$$X_i = \begin{cases} 1 & \text{si pile,} \\ 0 & \text{si face.} \end{cases}$$

- Modèle : $X_1, \dots, X_{1000}$ indépendants et de loi $\mathcal{B}(p)$ (p inconnu).
Le paramètre d'intérêt θ est $\theta = p$.
- On veut tester $H_0: p = 0.5$ contre $H_1: p \neq 0.5$.
- Statistique de test : $\hat{p} = \frac{X_1 + \dots + X_{1000}}{1000}$.
- Région de rejet : $|\hat{p} - 0.5| \geq S$.
(Situation où on décide H_1).
Idée : on accepte H_0 si $|\hat{p} - 0.5| \leq S$

En effet, si H_0 est vraie, $|\hat{p} - 0.5|$ a tendance à être plus petit que si c'est H_1 qui est vraie.

- Choix de S : pour avoir un test de niveau α (par exemple : $\alpha = 5\%$) (probabilité de se tromper quand H_0 vraie est $\leq \alpha$),
On veut que si $p = 0.5$, $P(|\hat{p} - 0.5| \geq s) \leq \alpha$.
On peut utiliser B-T: $P(|\hat{p} - 0.5| > s) \leq \frac{0.25}{10000 \cdot s^2}$ car:

- $\hat{p} = \frac{X_1 + \dots + X_n}{n}$
- $\mathbb{E}[X_1] = 0.5$
- $\text{Var}(X_1) = 0.5(1 - 0.5) = 0.25$
- $\mathbb{E}(\hat{p}) = 0.5$
- $\text{Var}(\hat{p}) = \frac{0.25}{10000}$

On choisit s de sorte que $\frac{0.25}{10000 \cdot s^2} \leq \alpha \Leftrightarrow s^2 \geq \frac{0.25}{10000 \cdot \alpha} \Leftrightarrow s \geq \frac{0.5}{10^2 \sqrt{\alpha}}$

si $\alpha = 1\% = 0.01$ $s \geq \frac{0.5}{10^2 \sqrt{10^{-2}}} = 0.05$

- Règle de décision:

- si $|\hat{p} - 0.5| < 0.05$ on décide H_0 (c-a-d $0.45 < \hat{p} < 0.55$)
- si $|\hat{p} - 0.5| \geq 0.05$ on décide H_1

B-T donne une majoration très grossière.

Mieux: théorème de limite centrale (l'an prochain)

Encore mieux: table exacte de lois

2. On se demande si l'abondance d'une espèce rare a diminué à tel endroit et à telle période, abondance moyenne d'individu de Z

On a 3 observations: (à ce lieu, à cette période, 3 temps)

Donnés résultats de Z_1, Z_2, Z_3 variables aléatoires.

- Modèle: Z_1, Z_2, Z_3 sont des v.a indépendantes de loi $P(\lambda)$ ou $\lambda = \mathbb{E}[Z_1]$ est l'abondance (on s'intéresse à $\theta = \lambda$)
- On veut tester: H_0 : " $\lambda = 2$ " contre H_1 : " $\lambda < 2$ "
- Statistique de test: $\frac{Z_1+Z_2+Z_3}{3} = \hat{\lambda}$ estimateur sans biais de λ
- Région de rejet: $\hat{\lambda} < s$ (En effet si H_1 est vraie, $\hat{\lambda}$ a tendance à être plus petit que si H_0 est vraie)
- Choix de s : pour avoir un test de niveau α
On veut que si $\lambda = 2$, $P\left(\frac{Z_1+Z_2+Z_3}{3} < s\right) \leq \alpha$
 Z_1, Z_2 et Z_3 indépendantes et de loi de poisson, donc $Z_1 + Z_2 + Z_3$ de loi de poisson.
Si $\lambda = 2$, $Z_1 + Z_2 + Z_3 \sim \mathcal{P}(6)$

$$P\left(\frac{Z_1 + Z_2 + Z_3}{3} < s\right) = P(Z_1 + Z_2 + Z_3 < 3s) = P(Z_1 + Z_2 + Z_3 \leq 3s - 1)$$

- Table de la loi de Poisson $\mathcal{P}(6)$: $Z \sim \mathcal{P}(6)$

k	0	1	2	3	4
$P(Z \leq k)$	.	.	.	.	.
	$P(Z = 0)$	$P(Z \leq 1)$	$P(Z \leq 2)$	$P(Z \leq 3)$	$P(Z \leq 4)$

- Règle de décision:

- Si $\frac{Z_1+Z_2+Z_3}{3} < \frac{4}{3}$ (c-a-d si $Z_1 + Z_2 + Z_3 < 4$) - On décide H_1 (on rejette H_0)
- Si $Z_1 + Z_2 + Z_3 \geq 4$ - On décide H_0

- Erreur de 2^{ème} espèce?

Probabilité de se tromper quand H_1 est vraie.

Si $\lambda < 2$, $P(Z_1 + Z_2 + Z_3 \geq 4)$ avec $Z_1 + Z_2 + Z_3 \sim \mathcal{P}(3\lambda)$ avec $3\lambda < 6$ fonction de la valeur de λ

- Puissance du test: probabilité d'avoir raison en décidant $H_1 = 1 - \text{erreur de 2^{ème} espèce}$.

- Fonction de répartition de F d'une v.a X (= CDF)

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0, 1] \\ t &\longmapsto F(t) = P(X \leq t). \end{aligned}$$

Si $X_1, \dots, X_n$ indépendantes et de même loi que X . Soit t un réel fixé.

$$F(t) = P(X \leq t) = \mathbb{E}[1_{X \leq t}].$$

Estimateur de $F(t)$?

$$\hat{F}(t) = \frac{1}{n} (1_{X_1 \leq t} + \dots + 1_{X_n \leq t}).$$