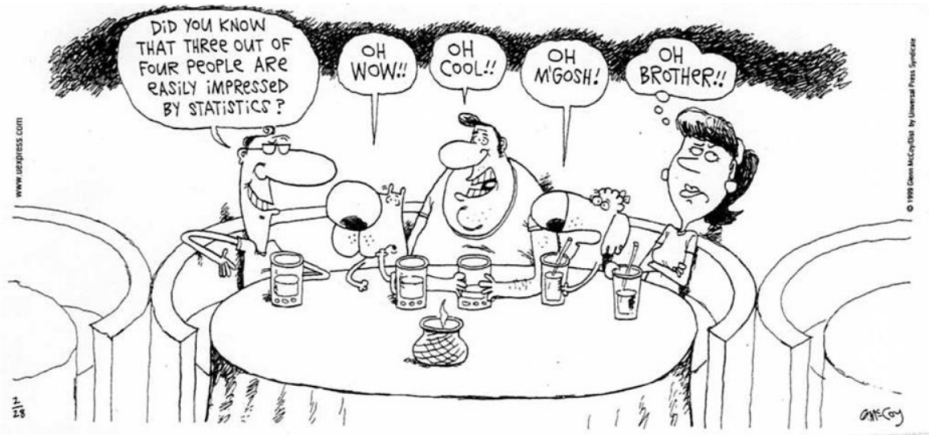
Estimation



Dans ce chapitre, on s'intéresse au problème de trouver la loi suivie par une variable aléatoire à partir de données statistiques. En pratique, on restreint la recherche à une famille de lois dépendant d'un paramètre noté θ, par exemple l'ensemble des lois de Poisson de paramètre θ, et on cherche à déterminer la valeur de ce paramètre.

On reproduit n fois l'expérience et on note X_1, X_2, …, X_n les variables aléatoires correspondant aux résultats de ces n expériences.

1 Premier exemple

On tire successivement et avec remise un jeton dans un sac opaque contenant des jetons blancs et des jetons noirs avec une proportion θ de jetons noirs, que l'on cherche à déterminer. C'est-à-dire que la probabilité de tirer un jeton noir est égale à θ. On note X la variable aléatoire égale à 1 si le jeton tiré est noir, et à 0 sinon. X suit une loi de Bernoulli de paramètre θ.

Pour cela on effectue 10 tirages successifs, et on note X_1, X_2, …X_{10} le résultat obtenu au 1^{er}, $2^{\text{ème}}$, … , $10^{\text{ème}}$ tirage.

On a $E\left(\dfrac{X_1+X_2+\cdots+X_{10}}{10}\right) = \dfrac{1}{10}E\left(X_1+X_2+\cdots+X_{10}\right) = \dfrac{1}{10}(\theta+\theta+\cdots+\theta) = \theta$

La variable aléatoire $T_{10} = \dfrac{X_1+X_2+\cdots+X_{10}}{10}$ est appelée estimateur de θ dans la mesure où elle a vocation à fournir une estimation de la valeur de θ.

2 Deuxième exemple

Un fabricant de pneus relève le nombre de kilomètres parcourus par cent pneus neufs d'un modèle donné avant la première crevaison et obtient les résultats suivants :

208	1685	554	1348	1037	226	1405	211	1305	44
1321	1421	141	680	3044	184	102	65	1016	1376
21	757	96	336	684	620	167	150	1265	764
468	438	289	345	577	507	910	1832	2030	208
455	2493	56	1499	1086	83	83	1332	88	1379
138	223	121	2816	288	740	616	53	2408	1610
60	1678	423	656	199	129	168	720	580	192
1112	632	458	697	2053	888	467	721	1865	41
2618	46	530	398	1572	119	364	229	142	632
356	770	136	235	700	504	1453	125	193	717

Il note X la variable aléatoire égale au nombre de kilomètres parcourus par un pneu et fait l'hypothèse que X suit une loi exponentielle de paramètre $\dfrac{1}{\theta}$. Par conséquent $E(X) = \dfrac{1}{\frac{1}{\theta}} = \theta$.
On note X_1, X_2, …, X_{100} les variables aléatoires égales au nombre de kilomètres parcourus par le 1er, le 2ème, … , 100ème pneu.
On a $E\left(\dfrac{X_1 + X_2 + \cdots + X_{100}}{100}\right) = \dfrac{1}{100} E\left(X_1 + X_2 + \cdots + X_{100}\right) = \dfrac{1}{100}(\theta + \theta + \cdots + \theta) = \theta$.
La variable aléatoire $\dfrac{X_1 + X_2 + \cdots + X_{100}}{100}$ est un estimateur de θ.

3 Cadre théorique

Définition : Estimateur.

Étant donné un n-échantillon de variables aléatoires $(X_1, X_2, \ldots, X_n)$ indépendantes et de même loi qu'une variable aléatoire X dépendant d'un paramètre θ, un estimateur de θ est une variable aléatoire de la forme $T_n = \varphi(X_1, X_2, \ldots, X_n)$.

Dans le premier exemple ci-dessus, $T_{10} = \dfrac{X_1 + X_2 + \cdots + X_{10}}{10}$ est un estimateur de la proportion θ de jetons noirs.

L'estimateur ne doit évidemment pas dépendre du paramètre θ !

Définition : Biais d'un estimateur.

Si T_n est un estimateur qui admet une espérance, on appelle biais de l'estimateur T_n le réel :

$$b_\theta(T_n) = E(T_n) - \theta$$

Si $E(T_n) = \theta$, soit $b_\theta(T_n) = 0$, on dit que l'estimateur est **sans biais**.

Exemple : Les deux estimateurs présentés en introduction sont des estimateurs sans biais.

Définition : Risque quadratique.

Si T_n^2 admet une espérance pour tout valeur de θ, on appelle risque quadratique de l'estimateur le réel :

$$r_\theta(T_n) = E((T_n - \theta)^2)$$

Plus le risque quadratique est faible, meilleure est la qualité de l'estimateur.

En développant le risque :

$$E((T_n - \theta)^2 = E(T_n^2) - 2\theta E(T_n) + \theta^2$$

Or :

$$b_\theta(T_n)^2 = E(T_n)^2 - 2\theta E(T_n) + \theta^2$$

On obtient donc :

Théorème.

Si T_n est un estimateur qui admet une espérance et une variance, alors :

$$r_\theta(T_n) = b_\theta(T_n)^2 + V(T_n)$$

Exemple : Revenons à la situation des jetons.
Nous avons vu que T_{10} est un estimateur sans biais de la proportion θ de jetons noirs.
Son risque quadratique est :
$r_\theta(T_{10}) = V(T_{10}) = V\left(\dfrac{X_1 + X_2 + \cdots + X_{10}}{10}\right) = \dfrac{1}{100}V(X_1 + X_2 + \cdots + X_{10}) = \dfrac{\theta(1-\theta)}{10}$.
Comparons ce résultat avec celui d'un autre estimateur :
X_1 est un estimateur de θ sans biais puisque $E(X_1) - \theta = 0$.
Son risque quadratique est $r_\theta(X_1) = V(X_1) = \theta(1-\theta)$
T_{10} a donc un risque quadratique dix fois plus faible que X_1 : c'est un estimateur plus fiable.

Remarque : En particulier, comme nous l'avons vu dans les exemples ci-dessus, lorsque l'estimateur est sans biais, le risque quadratique est égal à la variance de l'estimateur.

4 Estimation par intervalle de confiance

Intervalle de confiance

Soient U_n et V_n deux estimateurs. On dit que $[U_n; V_n]$ est un intervalle de confiance de θ au niveau de confiance $1-\alpha$ où $\alpha \in [0;1]$ si, pour tout θ, $P([U_n \leq \theta \leq V_n]) \geq 1 - \alpha$.

Exemple.
Cas d'un sondage. On interroge n personnes qui doivent se prononcer par Oui ou Non à un référendum. On suppose que la réponse de chaque personne suit est une variable aléatoire qui suit une loi de Bernoulli de paramètre p (Oui pour le "Succès," Non pour l'"Echec"). On interroge n personnes, on a donc un n-échantillon de variables aléatoires indépendantes $(X_1, \ldots, X_n)$ de même loi $\mathcal{B}(p)$.

On note $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. C'est un estimateur de la proportion de "Oui".

Alors $\overline{X_n}$ est un estimateur sans biais de p, et $V(\overline{X_n}) = \frac{p(1-p)}{n}$. Or $p(1-p) \leq \frac{1}{4}$ pour tout $p \in [0;1]$.

Appliquons l'inégalité de Bienaymé-Tchebytchev : $P(|\overline{X_n} - p| \geq \epsilon) \leq \frac{V(\overline{X_n})}{\epsilon^2} \leq \frac{1}{4n\epsilon^2}$

Cette inégalité équivaut à $P(\overline{X_n} - \epsilon \leq p \leq \overline{X_n} + \epsilon) \geq 1 - \frac{1}{4n\epsilon^2}$

Donc, en posant $U_n = \overline{X_n} - \epsilon$ et $V_n = \overline{X_n} + \epsilon$, l'intervalle $[U_n; V_n]$ est un intervalle de confiance de p au niveau de confiance $1-\alpha$, où $\alpha = \frac{1}{4n\epsilon^2}$.

Exemple 1 : n donné

On prend $n = 100$. Si on cherche un intervalle de confiance au niveau 0.96, il suffit d'avoir $1 - \frac{1}{4n\epsilon^2} = 0.96$, soit $\epsilon = \frac{1}{4}$. L'intervalle $[\overline{X_n} - \frac{1}{4}, \overline{X_n} + \frac{1}{4}]$ est un intervalle de confiance au niveau 0.96.

Exemple 2 : ϵ donné

L'intervalle précédent est de largeur 0.5. Généralement on souhaite contrôler cette largeur pour connaître la précision, on cherche donc plutôt n tel que $1 - \frac{1}{4n\epsilon^2} \geq 0.96$.

Si on choisi $\epsilon = 0.1$ on obtient $n \geq 625$.

Si on choisi $\epsilon = 0.01$ on obtient $n \geq 62500$.