

Probabilités, Statistiques, Combinatoire : TD8

Statistiques : estimation, intervalle de confiance

Exercice 8.1

Estimation de l'équilibre d'un jeu

Un certain jeu à deux joueurs peut se terminer par la victoire du joueur (A) ou de l'autre (B), ou par une partie nulle (N). Le jeu fait intervenir le hasard ; on suppose que le joueur A a probabilité p de gagner, le joueur B, probabilité q , et que la partie est nulle avec probabilité $1 - (p + q)$. On ne connaît ni p , ni q , ni $p - q$; on souhaite estimer ces paramètres.

On suppose que l'on a accès à une séquence de n résultats de parties, sous la forme d'un mot $W = W_1 \dots W_n \in \{A, B, N\}^n$.

1. Pour le résultat d'une partie, $t \in \{A, B, N\}$, on définit les variables aléatoires X , Y et Z , par $X(A) = 1$, $X(B) = X(N) = 0$; $Y(B) = 1$, $Y(A) = Y(N) = 0$; et $Z = X - Y$. Déterminer les lois de X , Y et Z , ainsi que, pour chacune d'elles, son espérance et sa variance, le tout en fonction de p et q . En particulier, montrer que $\mathbf{var}(X) \leq 1/4$ et $\mathbf{var}(Z) \leq 1$.
2. À partir de l'échantillon $W = (W_1, W_2, \dots, W_n)$, on définit, pour chaque i ($1 \leq i \leq n$), $X_i = X(W_i)$, $Z_i = Z(W_i)$, et enfin $S_n = (X_1 + \dots + X_n)/n$ et $T_n = (Z_1 + \dots + Z_n)/n$. Montrer que S_n est un estimateur sans biais de p , et que T_n est un estimateur sans biais de $p - q$.
3. Déterminer les variances de S_n et de T_n .
4. On note $\sigma_n = \sqrt{\mathbf{var}(S_n)}$ (σ_n est donc inconnu). On fixe un paramètre $\epsilon > 0$. Montrer qu'on a $\epsilon \geq (2\sqrt{n}\epsilon)\sigma_n$, et que l'événement $\{p - (2\sqrt{n}\epsilon)\sigma_n \leq S_n \leq p + (2\sqrt{n}\epsilon)\sigma_n\}$ est inclus dans l'événement $\{p - \epsilon \leq S_n \leq p + \epsilon\}$. En déduire que S_n est un estimateur convergent pour p (utiliser l'inégalité de Tchebycheff).
5. Adapter le calcul précédent à $\sigma'_n = \sqrt{\mathbf{var}(T_n)}$, et montrer que T_n est un estimateur convergent pour $p - q$.
6. Identifier, en fonction de n , une valeur a_n telle qu'on puisse affirmer que la probabilité de l'événement $\{p - a_n \leq S_n \leq p + a_n\}$ soit supérieure ou égale à 0.95, et une valeur b_n telle qu'on puisse affirmer que la probabilité de l'événement $\{p - q - b_n \leq T_n \leq p - q + b_n\}$. L'intervalle (aléatoire : il dépend de S_n) $[S_n - a_n, S_n + a_n]$ est appelé **intervalle de confiance, au niveau de risque 0.05, pour p** (de même pour l'intervalle $[T_n - b_n, T_n + b_n]$ pour $p - q$).
7. À quoi correspond, intuitivement, le fait que 0 se trouve dans un intervalle de confiance pour $p - q$? que 0 soit à gauche d'un tel intervalle ? à droite de l'intervalle ?
8. Pour différents jeux, on a relevé les résultats suivants (à chaque fois, on donne le nombre de victoires pour A, le nombre de victoires pour B, et le partie nulles ; chaque série de résultats porte *a priori* sur un jeu différent, donc des valeurs de p et q potentiellement différentes.

Pour chaque jeu, déterminer numériquement l'estimation et l'intervalle de confiance de la question précédente pour $p - q$. Pour quels jeux l'intervalle de confiance contient-il la valeur 0 ?

- Jeu 1 : 100 victoires pour A, 72 victoires pour B, 28 parties nulles.
- Jeu 2 : 1000 victoires pour A, 720 victoires pour B, 280 parties nulles.
- Jeu 3 : 138 victoires pour A, 237 victoires pour B, 0 parties nulles.