La méthode BIT pour le calcul des probabilités au hockey

par Alain Bonnier, D.Sc. (physique)

L'équipe qui gagne un match de hockey est celle qui compte le plus de buts ou s'en fait compter le moins ! L'affirmation semble triviale mais montre l'importance quand on veut prédire la performance d'une équipe, au hockey, d'arriver à estimer d'abord le nombre de buts qu'elle est susceptible de compter ou se faire compter lors d'un match. Cet estimé permettra ensuite d'évaluer les chances qu'aura cette équipe de gagner un match en particulier contre une équipe donnée. Et de là, d'extrapoler sa performance jusqu'en fin de saison.

La distribution des buts

Dans la Ligue Nationale de Hockey, on constate que le rendement d'une équipe présente des différences significatives selon qu'elle joue à domicile ou à l'extérieur. Ce qui justifie qu'on analyse séparément les matches joués dans l'une ou l'autre de ces situations.

Si on prend l'ensemble des N matches joués à domicile (ou à l'extérieur) par une équipe au cours d'une saison et qu'on compte le nombre de matches où elle a marqué k buts (k = 0, 1, 2, 3, etc., en excluant pour l'instant ceux marqués en prolongation), on forme alors un tableau des fréquences f_D. du nombre de buts comptés à domicile. Ces fréquences f_D divisées par N nous donnent la distribution de probabilité empirique $p_{D,emp}$ des buts comptés à domicile. Cette distribution permet un premier estimé des chances qu'a cette équipe de compter k buts au cours des 60 premières minutes d'un match. Pour le Canadien, par exemple, le tableau de cette distribution se présente ainsi pour la saison 96-97, après les N = 38 matches joués à domicile (entre le 4 octobre 1996 et le 2 avril 1997) :

Tableau 1

k	f_D	$p_{D,emp}$ %	$p_{D,thé}$ %
0	2	5,2	4,0
1	5	13,2	13,0
2	6	15,8	20,7
3	11	28,9	22,2
4	6	15,8	17,9
5	2	5,2	11,5
6	5	13,2	6,1
7	0	0,0	2,8
8	1	2,6	1,1
9	0	0,0	0,4
10	0	0,0	0,1

En calculant ensuite la moyenne μ_D des buts comptés à domicile

$$\mu_D = \frac{1}{N}\sum_{k=0}^{20} k f_{D,k} \qquad \text{(Eq. 1)}$$

ainsi que l'écart-type σ_D,

$$\sigma_D = \sqrt{\frac{1}{N}\sum_{k=0}^{20}(k-\mu_D)^2 f_{D,k}} \qquad \text{(Eq. 2)}$$

on peut ensuite comparer cette distribution empirique avec différentes distributions connues (distribution normale, binômiale, géométrique, hypergéométrique, de Poisson, de Pascal, de Fisher, de Student, etc.) pour voir si l'une d'elles s'en approche. La comparaison se fait généralement en calculant, dans chaque cas, l'erreur quadratique moyenne entre une distribution donnée et la distribution empirique. Pour le Canadien en 96-97, par exemple, $\mu_D = 3{,}21$ et $\sigma_D = 1{,}82$.

Pour décrire le nombre de buts comptés par matches au hockey, c'est la distribution de Poisson qui présente généralement la plus faible erreur quadratique moyenne (de l'ordre de 5% après 40 matches). Si on emploie cette distribution, la probabilité $p(\mu, k)$ qu'une équipe (ayant une moyenne offensive μ) compte k buts en l'espace de 60 minutes s'évalue à partir de la relation:

$$p(\mu, k) = \mu^k / k\,!\ \exp(-\mu) \qquad \text{(Eq. 3)}$$

où le symbole k ! représente la factorielle $k\cdot(k-1)\cdot(k-2)...3\cdot2\cdot1$ et où $\exp(-\mu)$ est la fonction exponentielle « e exposant $-\mu$ », avec e = 2,71828...

Une des propriétés de cette distribution est que la moyenne μ est égale à la variance σ^2. On peut voir, dans le cas du Canadien, que la variance observée σ_D^2 est égale à 3,31 alors que la moyenne observée $\mu_D = 3{,}21$. La différence entre les deux est d'à peine 3 %. J'ai présenté également, dans la dernière colonne du Tableau 1, la probabilité théorique $p_{D,thé}$ qu'on obtiendrait si on évaluait la probabilité de compter k buts à partir de l'Eq.3. On peut constater que les valeurs obtenues suivent de très près la probabilité empirique $p_{D,emp}$. C'est donc cette distribution qui a été retenue pour les fins du calcul.

L'espérance offensive

La moyenne μ des buts comptés, calculée à partir des N derniers matches joués, n'est cependant pas le paramètre le plus significatif pour déterminer l'espérance offensive λ (ou nombre probable de buts qu'une équipe va compter) lors d'un match donné. Cette espérance offensive λ est fonction bien sûr de la performance offensive passée d'une équipe (ce qu'exprime la moyenne de buts comptés) mais aussi de la performance défensive passée de l'équipe adverse. De plus, comme la performance d'une équipe varie au cours d'une saison, les résultats des matches les plus récents sont beaucoup plus significatifs que ceux des matches antérieurs. Pour tenir compte de ces facteurs, il est donc préférable de remplacer la moyenne μ dans l'Eq.3 par l'espérance offensive λ qui sera calculée à partir d'une analyse de régression multiple sur des moyennes exponentielles de buts.

Le problème se réduit donc à évaluer les espérances offensives λ_D et λ_E des équipes jouant à *D*omicile et à l'*E*xtérieur lors d'un match donné. Connaissant ces espérances, les probabilités de victoire ou de match nul se déduiront ensuite de la distribution de Poisson.

Par « moyenne exponentielle », on entend une moyenne où le poids relatif des matches les plus récents est supérieur à ceux antérieurs. La moyenne exponentielle b_N après N matches, se calcule par itération à partir de l'équation

$$b_N = b_{N-1} + (k_N - b_{N-1})/N_C \qquad \text{(Eq. 4)}$$

où b_{N-1} est la moyenne exponentielle après N–1 matches, k_N le nombre de buts comptés lors du Nième match et N_C est le nombre caractéristique de matches. Ce nombre N_C caractérise le taux de décroissance des poids relatifs en allant vers le passé. Il est obtenu empiriquement en cherchant la valeur de N_C qui minimise l'erreur quadratique moyenne dans l'analyse de régression linéaire multiple, décrite ci-après. Pour les séquences observées, la valeur de N_C se situe généralement autour de 20.

On suppose ensuite que l'espérance offensive λ_D de l'équipe D (celle qui joue à Domicile) dépend d'une certaine façon de la moyenne exponentielle b_{PDD} des buts Pour l'équipe D quand elle joue à Domicile, de celle b_{PDE} des buts Pour l'équipe D quand elle joue à l'Extérieur et de celles b_{CED} et b_{CEE} des buts Contre l'équipe E quand elle joue à Domicile et à l'Extérieur respectivement.

De même, on peut supposer que l'espérance offensive λ_E de l'équipe E (celle qui joue à l'Extérieur) doit dépendre des moyennes exponentielles b_{PED}, b_{PEE}, b_{CDD} et b_{CDE}. On peut vérifier si ces hypothèses ont du sens en regardant si les coefficients de corrélation entre ces espérances et ces moyennes sont positives. Pour ce faire, on suppose dans un premier temps, à défaut d'indication contraire, que la relation entre les espérances et les différentes moyennes exponentielles est linéaire. (Cette dernière hypothèse a été modifiée dans les versions subséquentes du modèle. Une relation polynômiale du 3ème degré semble donner une meilleure précision pour les équipes qui se situent aux extrêmes par rapport à la moyenne de la Ligue.)

Ainsi, dans l'hypothèse d'une relation linéaire,

$$\lambda_D = c_{D0} + c_{D1}\, b_{PDD} + c_{D2}\, b_{PDE} + c_{D3}\, b_{CEE} + c_{D4}\, b_{CED} \qquad \text{(Eq. 5)}$$

et

$$\lambda_E = c_{E0} + c_{E1}\, b_{PEE} + c_{E2}\, b_{PED} + c_{E3}\, b_{CDD} + c_{E4}\, b_{CDE} \qquad \text{(Eq. 6)}$$

où les différentes valeurs de c_D et c_E sont celles qui minimisent l'erreur quadratique moyenne sur λ_D et λ_E. Pour la saison 96-97 de la Ligue Nationale de Hockey, par exemple, on obtient

$$\lambda_D = 1{,}378 + 0{,}765\, b_{PDD} + 0{,}483\, b_{PDE} + 0{,}761\, b_{CEE} + 0{,}261\, b_{CED} \qquad \text{(Eq. 7)}$$

et

$$\lambda_E = 0{,}645 + 0{,}782\, b_{PEE} + 0{,}399\, b_{PED} + 0{,}391\, b_{CDD} + 0{,}694\, b_{CDE}. \qquad \text{(Eq. 8)}$$

La valeur des coefficients c_{Di} et c_{Ei} nous donne une idée de l'importance relative de ces différents facteurs. Ainsi, on peut voir que pour déterminer le nombre probable de buts comptés par une équipe jouant à domicile, par exemple, la moyenne b_{PDD} de buts comptés à domicile avant un match semble jouer un rôle 66% plus important que la moyenne b_{PDE} de buts comptés à

l'extérieur par cette même équipe. Par contre la moyenne b_{CED} des buts comptés par l'équipe visiteuse lorsqu'elle jouait à domicile, joue un rôle trois fois moins important que la moyenne $b_{PDD.}$

La simulation d'un match

Connaissant les espérances offensives λ_D et λ_E de deux équipes qui se rencontrent, on peut calculer les probabilités P_D et P_E que les équipes D ou E gagnent ainsi que la probabilité P_N que le résultat soit nul, à partir des relations

$$P_N = P_{NR} \times P_{NS} \qquad \text{(Eq. 9)}$$

où

$$P_{NR} = \sum_{k=0}^{\infty} p(\lambda_D, k) \cdot p(\lambda_E, k) \qquad \text{(Eq. 10)}$$

est la probabilité d'obtenir un résultat nul après les 60 minutes de temps *R*égulier, et où

$$P_{NS} = p(\lambda_D/12,\ 0) \cdot p(\lambda_E/12,\ 0) \qquad \text{(Eq. 11)}$$

la probabilité que le résultat soit nul après les 5 minutes de temps *S*upplémentaire.

De même, la probabilité que l'équipe D gagne est donnée par la relation

$$P_D = \sum_{k_D=1}^{\infty} \left[p(\lambda_D, k_D) \cdot \sum_{k_E=0}^{k_D-1} p(\lambda_E, k_E) \right] + P_{NR} \cdot \left[1 - P_{NS}\right] \cdot \left[\lambda_D \Big/ (\lambda_D + \lambda_E) \right] . \qquad \text{(Eq. 12)}$$

Finalement, la probabilité que l'équipe E gagne se déduit des deux autres par la relation

$$P_E = 1 - P_N - P_D . \qquad \text{(Eq. 13)}$$

La méthode stochastique

Mais on peut aussi évaluer le nombre probable de buts comptés par les équipes D et E en générant deux nombres aléatoires a_D et a_E obéissant à la distribution de Poisson d'espérance λ_D et λ_E respectivement.

Pour générer un nombre aléatoire a obéissant à la distribution de Poisson, il suffit de générer un nombre aléatoire r de distribution uniforme, compris entre 0 et 1 (tel que donné généralement par la fonction RND de la plupart des langages de programmation), et de trouver la première valeur de $a = k$ (en commençant par $k = 0$) pour laquelle la condition suivante est satisfaite

$$\sum_{k=0}^{a} p(\lambda, k) \geq r \ . \qquad \text{(Eq. 14)}$$

Une fois ces deux nombres a_D et a_E obtenus, on les compare entre eux. Si $a_D > a_E$, on compte une victoire pour l'équipe *D*. Si $a_D < a_E$, on enregistre une victoire pour l'équipe *E*. Et si Si $a_D = a_E$, on génère deux autres nombres aléatoires pour simuler la période de prolongation et voir si l'égalité persiste. En simulant ainsi un match *n* fois et en comptabilisant les résultats on arrive à une précision de l'ordre de $100/\sqrt{n}$, en pourcent. Cette façon de calculer les probabilités s'appelle «méthode stochastique». Pour n = 10 000, par exemple, on obtient une précision d'environ 1% dans l'évaluation de la probabilité du résultat d'un match.

Cependant les moyennes de buts μ, servant à calculer les espérances λ, sont obtenues à partir d'un échantillonnage restreint de *N* matches. Elles sont donc entachées d'une certaine incertitude de l'ordre de l'écart-type σ. Ce qui affecte la confiance que l'on peut avoir dans le calcul de l'espérance. On peut traduire cette incertitude en générant une nouvelle espérance λ^* suivant une distribution normale centrée sur λ et d'écart-type σ à partir de la relation

$$\lambda^* = \lambda + \sigma\sqrt{-2\ln r_1} \cdot \sin(2\pi r_2) \qquad \text{(Eq. 15)}$$

où r_1 et r_2 sont deux nombres aléatoires de distribution uniforme entre 0 et 1. C'est cette espérance λ^* qu'on utilise alors à la place de λ pour générer les nombres de buts aléatoires *a*.

L'avantage de la méthode stochastique par rapport à la méthode directe du calcul des probabilités, apparaît quand on veut voir, par exemple, quelles sont les chances d'une équipe de se qualifier pour les séries éliminatoires ou gagner le championnat de sa division, etc. Quand il reste 10, 20 ou 40 matches à jouer en saison, le calcul direct devient monstrueusement compliqué puisqu'il faut alors multiplier entre elles les probabilités de chaque match, dans toutes les combinaisons possibles. La méthode stochastique permet d'éviter cette complication. Il suffit de simuler tous les matches qui restent à jouer jusqu'à la fin de la saison par l'ensemble des équipes de la ligue et à comptabiliser les différentes issues en fin de saison. En répétant de la sorte *n* saisons et en divisant les totaux obtenus par *n*, on arrive ainsi à évaluer toutes les probabilités qui nous intéressent avec une précision de l'ordre de $100/\sqrt{n}$, en pourcent.

Si le calcul de probabilité est réaliste, on peut s'amuser à retenir les issues dont la probabilité d'occurrence dépasse 95% (ou celles inférieures à 5%) pour en faire des prédictions qui se réaliseront dans plus de 95% des cas. On se trompera donc moins d'une fois sur 20 en moyenne.

Et si vous publiez ensuite ces prédictions dans les médias, cette précision devrait être suffisante pour vous bâtir une solide réputation de « prophète sportif »...