La Statistique et les Classements

Les données issues de classement sont présentes partout : dans les compétitions sportives, sur les sites de réservations en ligne ou sur Parcoursup, la plateforme de préinscription en première année de l'enseignement supérieur en France. Voici quelques projets de recherche actuels en statistique sur le thème des classements.

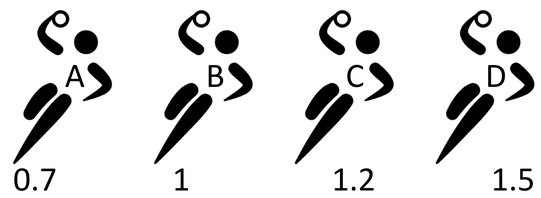
CLASSEMENT DES ÉQUIPES EN HANDBALL

Le **modèle de Bradley-Terry**, développé dans les années 1950 par les statisticiens Ralph Bradley et Milton Terry, est une méthode utilisée pour évaluer la force relative ou le classement des éléments dans une compétition, comme des équipes sportives, des produits ou des candidats politiques.

Expliquons son fonctionnement sur le **classement de 4 équipes en handball.** Les équipes A et B sont des équipes d'amateurs, avec un paramètre de force de 0.7 et 1 respectivement. Les équipes C et D sont des professionnels, avec un paramètre de force de 1.2 et 1.5 respectivement. Lorsque deux équipes s'affrontent, celle avec la qualité la plus élevée a plus de chances de gagner, mais même l'élément le moins fort a une chance de gagner, surtout si la différence est petite, ce qui augmenterait son paramètre de force pour les prochains matchs !

Dans un championnat, les 4 équipes se rencontrent. Voici les résultats dans le tableau ci-dessous.

	B	C	D
A	17-23	17-35	18-41
B		28-26	26-35
C			31-33



	A	B	C	D
Avant	0.7	1	1.2	1.5
Après	0.6	0.9	1.2	1.7

On met alors à jour les paramètres de force, comme dans la deuxième ligne, et le classement correspondant : l'équipe B est alors deuxième.

Le modèle de Bradley-Terry est largement utilisé dans de nombreux domaines, notamment les sports, le marketing, les élections et la recherche sociale. Une des grandes forces de ce modèle est sa simplicité et son adaptation à différents contextes.

Cependant, ce modèle repose sur certaines **hypothèses** qui peuvent être irréalistes en pratique, comme l'indépendance des résultats des matchs (l'équipe B s'est améliorée parce qu'elle a fait un bon recrutement) et la stabilité des paramètres de force au fil du temps, ou encore la transitivité des résultats (on évite les cycles A gagne contre B, B gagne contre C, et C gagne contre A).

La modélisation repose sur la comparaison par paire: par exemple, si on note $\mathbb{P}(i > j)$ la probabilité que le score de l'équipe i soit supérieur au score de l'équipe j,

$$\mathbb{P}(1 > 2) = \frac{p_1}{p_1 + p_2} = \frac{0.7}{1 + 0.7} = 0.41.$$

Pour voir l'évolution des paramètres de force, on calcule la **vraisemblance** associée à ce modèle après les matchs : si w_{ij} est le score de l'équipe i dans le match contre l'équipe j, c'est $\prod_{i \neq j} (\mathbb{P}(i > j))^{w_{ij}}$.

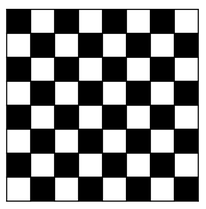
On met alors les paramètres à jour successivement : $p_1 = \frac{17 \times \frac{1.1}{1.1+0.5} + 17 \times \frac{1.5}{1.5+0.5} + 18 \times \frac{2.1}{2.1+0.5}}{\frac{23}{1.1+0.5} + \frac{35}{1.5+0.5} + \frac{41}{2.1+0.5}} = 33, p_2 = 53, p_3 = 76, p_4 = 106.$

On renormalise : $p_1 = \frac{p_1}{(p_1 p_2 p_3 p_4)^{1/4}} = 0.54, p_2 = 0.87, p_3 = 1.25, p_4 = 1.72,$

Et on itère jusqu'à convergence, pour obtenir $p_1 = 0.6, p_2 = 0.9, p_3 = 1.2, p_4 = 1.7.$

C'est la méthode de Dijkstra. On peut aussi utiliser la régression logistique.

CLASSEMENT DES JOUEURS D'ÉCHEC



Le **classement Elo** est une méthode (dérivée du modèle de Bradley-Terry) utilisée pour classer des joueurs dans des compétitions comme les échecs, les jeux vidéo, et même les sports. Inventé dans les années 1950 par Arpad Elo, un physicien et joueur d'échecs, ce système est très reconnu pour sa précision et sa simplicité.

Dans le cas **des échecs,** chaque joueur commence avec un score de base, 1200 points. On cherche ensuite à prédire le résultat des parties : étant donnés deux joueurs A et B de score 1600 et 1400 respectivement, l'**espérance de gain du joueur A** est donnée par $f_N(1600 - 1400) = 0.76$: le joueur A marquera en moyenne 0.76 point par partie contre le joueur B, où f_N est la fonction de répartition de la loi normale (voir l'encadré). Néanmoins, on ne sait pas comment mettre à jour le classement des joueurs en fonction des résultats d'une partie.

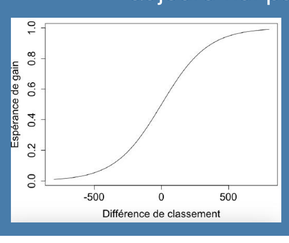
Avec la **méthode incrémentale**, on augmente ou diminue le classement selon qu'un joueur gagne ou perde un match, avec la formule suivante : $NS = AS + K \times (SO - EG)$ où K est un facteur qui détermine la sensibilité du score aux résultats du match, SO est le score obtenu, 1 pour une victoire, 0.5 pour un match nul, et 0 pour une défaite, et EG est l'espérance du gain. Dans l'exemple on obtient :

- Si A gagne, son nouveau score sera : 1600 + 20 × (1 - 0.76) = 1600 + 4.8 ≈ 1605.
- Si B gagne, son nouveau score sera : 1400 + 20 × (1 - 0.24) = 1400 + 15.2 ≈ 1415.

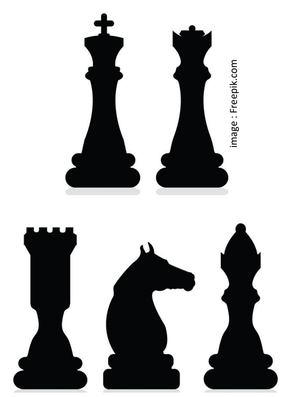
Le classement Elo, contrairement au modèle de Bradley-Terry, s'ajuste **dynamiquement** après chaque match, ce qui le rend très réactif aux performances actuelles des joueurs : on suit précisément l'évolution du niveau des joueurs au fil du temps. De plus, le système Elo est simple à comprendre et à appliquer : c'est un outil populaire dans de nombreux domaines compétitifs, utilisé encore aujourd'hui par la Fédération internationale des échecs.

Si la méthode fonctionne bien pour les joueurs dont le classement est établi, elle n'est pas adaptée aux nouveaux joueurs. Pendant leurs premiers matchs, les scores des nouveaux joueurs peuvent être plus volatils. Le facteur K peut être temporairement ajusté pour refléter plus rapidement leur véritable niveau. Cela permet d'avoir des compétitions plus équilibrées et justes.

La fonction f_N utilisée doit être une fonction croissante (plus la différence de points est grande, plus on a de chance de gagner) et bornée entre 0 et 1 (il s'agit de probabilités). En pratique, la fonction logistique est utilisée aujourd'hui par la Fédération internationale des échecs.



Ecart de classement	Espérance de gain
800	0.998
500	0.961
300	0.856
200	0.76
100	0.638
0	0.5



ALGORITHME DE TRI PAR INSERTION, RANGS ET MODÈLE PROBABILISTE

L'algorithme de tri par insertion est régulièrement utilisé, par exemple pour trier les cartes en jouant au tarot. Il est optimal quand le nombre d'objets à trier est inférieur à 10.
Il s'agit de considérer chaque élément et de l'insérer à la bonne place parmi les cartes déjà triées. Pour trouver sa place, il faut comparer l'élément à ceux déjà triés, deux à deux.
Par exemple, si on trie l'ensemble 8♥, 10♠, 7♠, V♥, 10♥, 9♠, on laisse le 8♥ et le 10♠ à leur place, puis on déplace le 7♠ pour le mettre avant le 10♠, et on fait cela pour tous les éléments. Cela donne, successivement :

8♥, 7♠, 10♠, V♥, 10♥, 9♠
8♥, V♥, 7♠, 10♠, 10♥, 9♠
8♥, 10♥, V♥, 7♠, 10♠, 9♠
8♥, 10♥, V♥, 7♠, 9♠, 10♠

Le rang d'un élément peut être vu comme le résultat d'un algorithme de tri, ici le tri par insertion.

Cependant, on peut faire des **erreurs sur un classement** : par exemple, on a demandé en 2012 à des personnes de classer les pays suivants : 1. France, 2. Allemagne, 3. Brésil et 4. Italie, en fonction du nombre de victoires à la coupe du monde de football. On observe que 69% des réponses comporte au moins une erreur de classement, qui s'interprète comme une (ou plusieurs) erreur(s) de comparaison dans l'algorithme de tri.
Si on souhaite minimiser les erreurs, il faut alors minimiser les comparaisons entre paires d'objets. On peut alors définir un **modèle probabiliste** qui représente ces erreurs, et comprendre les erreurs de classement.

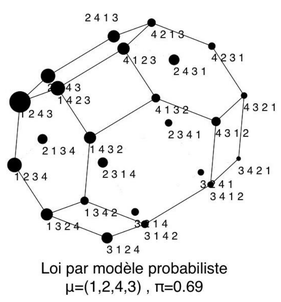
On représente une erreur dans une comparaison de paires par une **loi de Bernoulli** de paramètre $1 - \pi$ par rapport à une vérité notée μ. Si on part de l'ordre initial y d'entrée dans l'algorithme de tri, la probabilité d'aboutir à l'ordre x est $\mathbb{P}(x \mid y; \mu, \pi) = \pi^{B(x,y,\mu)}(1 - \pi)^{M(x,y,\mu)}$ avec

$B(x, y, \mu)$ le nombre total de bonnes comparaisons de paires,
$M(x, y, \mu)$ le nombre total de mauvaises comparaisons de paires.

On obtient finalement la loi de probabilité suivante, avec m le nombre d'objets à classer (dans la coupe du monde de football, m=4)

$$\mathbb{P}(x; \mu, \pi) = \frac{1}{m!} \sum_{y} \mathbb{P}(x \mid y; \mu, \pi).$$

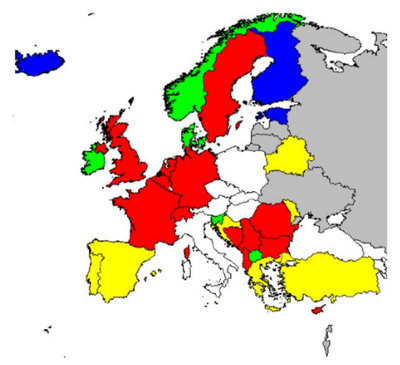
On représente cette loi probabilité dans le cas de l'exemple de la coupe du monde de football dans la figure ci-jointe, où la taille des points est proportionnelle à la probabilité. Même si le bon classement est le plus fréquent, on voit la répartition des erreurs.



Loi par modèle probabiliste
μ=(1,2,4,3) , π=0.69

CLASSEMENT À L'EUROVISION : DES RESSEMBLANCES AU SEIN DES PAYS ?

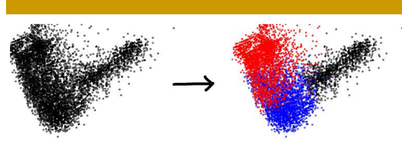
Dans le concours Eurovision de la chanson, qui est le plus grand concours musical au monde, les 24 pays participants classent les 10 premières chansons étrangères par ordre de préférence. Ce **classement est dit incomplet ou partiel** : on ne connait que les rangs 1 à 10.
On voudrait savoir s'il existe des **groupes de pays ayant des goûts similaires.** Le problème, c'est que traiter ces données est une tâche compliquée : on a seulement une information partielle sur les notes.
On peut généraliser la modélisation probabiliste de l'algorithme de tri par insertion au cas des données partielles, en les considérant comme des **données manquantes**.
On cherche à construire des groupes de pays, donc on ajoute une représentation en groupes via un algorithme de classification sur cette modélisation probabiliste. Sur les données de l'Eurovision de 2007 à 2012, on obtient cinq groupes de pays qui ont des préférences musicales similaires, représentés sur la carte : en rouge les pays d'Europe de l'ouest, en jaune les pays méditerranéens, en bleu les pays du nord, en gris les pays d'Europe de l'est et en vert un groupe plus difficile à analyser.



Cela suggère qu'il pourrait y avoir des alliances géographiques entre certains pays dans leurs choix de chansons à l'Eurovision. Par exemple, le groupe gris a souvent bien classé l'Ukraine et la Russie.

Notons que les données avec des rangs partiels (classement des 10 premières chansons, ou podium des 3 premiers athlètes en compétition sportive par exemple), et de rang multivariées (ici plusieurs années, mais par exemple sur une enquête sur des vacances, on note la piscine et la qualité des desserts) sont très courantes, et les groupes permettent de construire des sous-populations homogènes, sur lesquelles on peut facilement faire des conclusions globales.

On cherche souvent à faire des groupes en statistique. Considérer que les données sont homogènes est souvent une hypothèse forte. A la place on considère des **données hétérogènes** organisées en **groupes de données homogènes.** Dans la figure ci-contre, on représente par un point chaque individu.



A gauche, le nuage de points, aucune structure n'apparaît. A droite, on a colorié chaque groupe : il y a en fait 3 profils, chaque individu appartient à un seul groupe et peut être vu comme une variation du profil central.

SFdS

La Société Française de Statistique assure la promotion de la statistique dans toutes ses composantes : RECHERCHE, ENSEIGNEMENT, APPLICATIONS.

Retrouvez sur notre site nos activités **: organisation d'évènements (conférences, soirées débats,..) , publications scientifiques, formation.**

Scannez moi pour télécharger ce poster