|  |
| --- |
| Ecole Supérieur d’Ingénieur Léonard de Vinci |
| Le gabarit des basketteurs et leurs performances |
| Deuxième année de p Projet de statistiques |
|  |
|  |
|  |
|  |
| LE YHUELIC Noé, LIPANI BIANCHI Diego, MARCELINO Auriane |

Table des matières

[INTRODUCTION 3](#_Toc132383188)

[LE RECUEIL DES DONNÉES 4](#_Toc132383189)

[ÉTUDE DES DONNÉES 5](#_Toc132383190)

[BIBLIOGRAPHIE 7](#_Toc132383191)

# INTRODUCTION

Au cours de la deuxième année de classe préparatoire au cycle ingénieur, le module introduction aux statistiques permet aux élèves de réaliser un projet de statistiques visant à appliquer l’ensemble des méthodes vues en classe sur un jeu de données choisi afin de répondre à une problématique.

Depuis sa création en 1891 par un professeur de sport canadien, le basket-ball a connu un franc succès particulièrement aux états unis et est, de nos jours, l’un des sports les plus populaires dans le monde entier. Le basket-ball est devenu un sport professionnel populaire avec des ligues comme la NBA en Amérique du Nord.

Les joueurs professionnels de basket sont réputés pour leur habilité, leur force et leur agilité et sont considérés comme certains des meilleurs athlètes du monde.

En plus d’être un sport divertissant et compétitif, le basket est un sport qui permet de développer son esprit d’équipe et promeut des valeurs telles que la coopération et le respect mutuel. Il est pratiqué par des millions de personnes de tout âge dans le monde entier en loisir et en compétition.

Les clichés disent qu’il faut être grand pour être fort au basket-ball. Nous pouvons en effet constater que la très grande majorité des basketteurs (si ce n’est tous) sont très grands.

L’étude statistique que nous allons vous présenter porte sur l’analyse de la relation entre le gabarit des joueurs professionnels en NBA et leurs performances sur le terrain.

# LE RECUEIL DES DONNÉES

Notre étude consistera à voir si le gabarit des joueurs influe sur leur niveau au basket. Nous avons donc essayé de comprendre ce qui selon nous définissait le mieux ces 2 concepts et de trouver des bases de données intéressantes.

C’est sur Kaggle, une plateforme que nous avons jugée fiable car elle a la confiance de certaines des plus grandes entreprises de science des données du monde, telles que Walmart ou Facebook, que nous avons trouvé ce que nous cherchions. En effet il y avait une base de données indiquant le nom, prénom, âge, et toutes les informations relatives aux performances de chaque joueur de NBA pour chaque saison depuis 1996 (nombre moyen de points par match, nombre moyen de rebonds par match etc…). Nous avons donc pu définir le gabarit en fonction de 2 variables : la taille et le poids. Ces 2 variables sont des variables numériques/quantitatives et continues qui définissent tout ce que nous avons besoin de savoir sur une personne concernant son gabarit. Pour décrire le niveau, nous nous sommes concentrés sur 3 paramètres : le nombre moyen de points par match, le nombre moyen de rebonds par match (un rebond étant le fait de capter la balle après que la balle ait rebondi sur la planche ou l’arceau du panier de basket à la suite d’un tir manqué de l’adversaire), et le nombre moyen d’assists (passe décisive) par match. Ces 3 variables sont aussi numériques et continues. À la suite de la définition de nos variables, vu que nous voulions analyser le basket actuel pour découvrir le “meilleur gabarit actuel au basket” nous avons choisi simplement d’analyser la dernière saison (2021/2022) pour pouvoir vraiment trouver ce qui définit un bon joueur de nos jours.

# ÉTUDE DES DONNÉES

Vu que nous nous concentrons sur une seule saison présente dans notre base de données, nous avons commencé par créer un nouveau fichier avec toutes ces données, que nous avons mises sur R.

Grâce au code suivant, nous avons également fait une vérification des données pour être sûr qu’il n’en manquait aucune et ainsi avoir des données fiables et robustes.

Une image contenant texte

Description générée automatiquement

Nous pouvons grâce au résultat suivant qu’il n’en manque pas :

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquementNous avons ensuite fait un profil moyen d’un joueur de NBA, c'est-à-dire les tendances globales concernant nos variables d’intérêt en faisant leur moyenne, pour avoir un ordre d’idée de ce joueur moyen avec ce code :

Et voici les résultats :

Une image contenant texte

Description générée automatiquement

Celui-ci mesure donc environ 198 cm pour 97 kg, il marque un peu plus de 8 points par match, réalise près de 3,5 rebonds par match ainsi qu’un peu moins de 2 assists.

Puis, nous avons remarqué sur R, grâce à un nuage de points et un coefficient de corrélation que la relation entre la taille et le poids se rapproche fortement d’une relation linéaire car ce coefficient est proche de 1 :

Une image contenant graphique

Description générée automatiquement



Nous allons donc considérer dans notre étude que la taille et le poids sont semblables et nous n’allons donc garder que la taille pour effectuer nos tests, en supposant que les résultats trouvés s’appliquent également au poids.

Pour l’étude de nos variables, nous avons commencé par calculer les coefficients de corrélation de chaque variable définissant le niveau, en fonction de la taille :

Une image contenant texte, intérieur, capture d’écran

Description générée automatiquement

Une image contenant logo

Description générée automatiquement

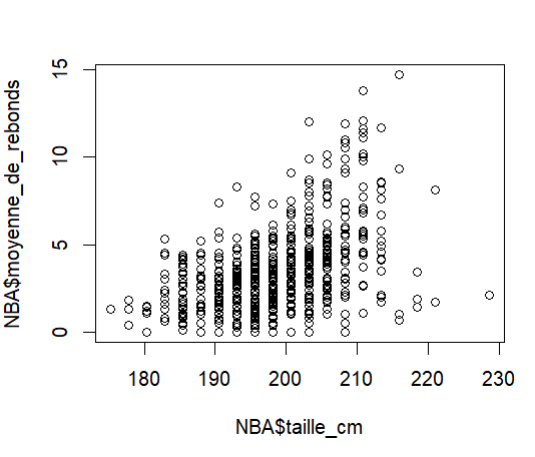


Une image contenant texte

Description générée automatiquement

A première vue, aucune variable de niveau ne forme une relation linéaire avec la taille, mais nous pouvons noter que c’est le nombre moyen de rebonds par matchs qui s’en rapproche le plus. Ensuite, nous avons effectué des nuages de points avec les mêmes variables, pour avoir une idée de la répartition de chaque observation concernant ces variables par rapport à la taille :

Une image contenant graphique

Description générée automatiquementUne image contenant graphique

Description générée automatiquement

Pour l’instant nous observons que les valeurs obtenues sont très espacées et ne paraissent pas suivre une tendance particulière, pour chaque variable la plupart des observations se trouvent entre 180cm et 215cm ce qui est très large.

Il faut donc maintenant que nous sachions quelle loi semblent suivre nos variables. Nous avons d’abord effectué des tests de Shapiro-Wilk sur R pour voir si nos variables suivent une loi normale :

Une image contenant texte

Description générée automatiquement

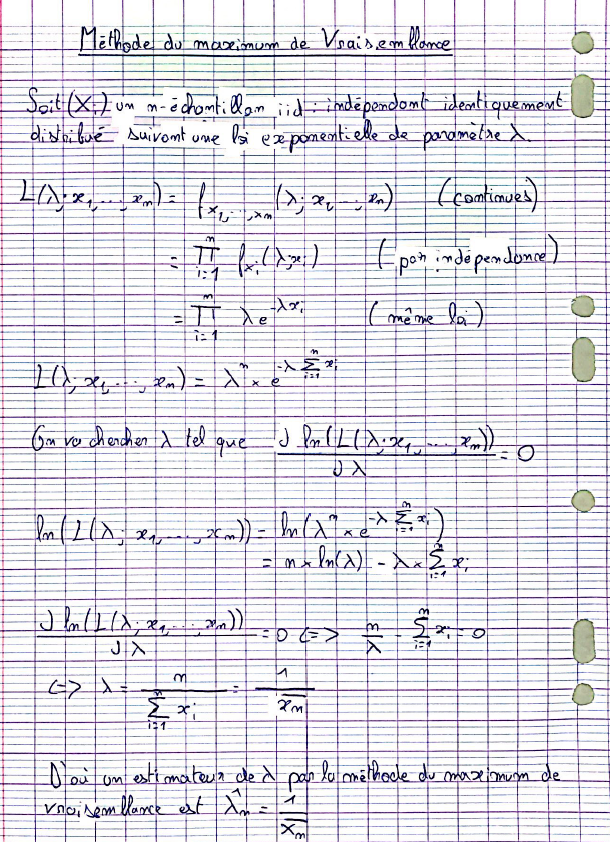
Une image contenant texte

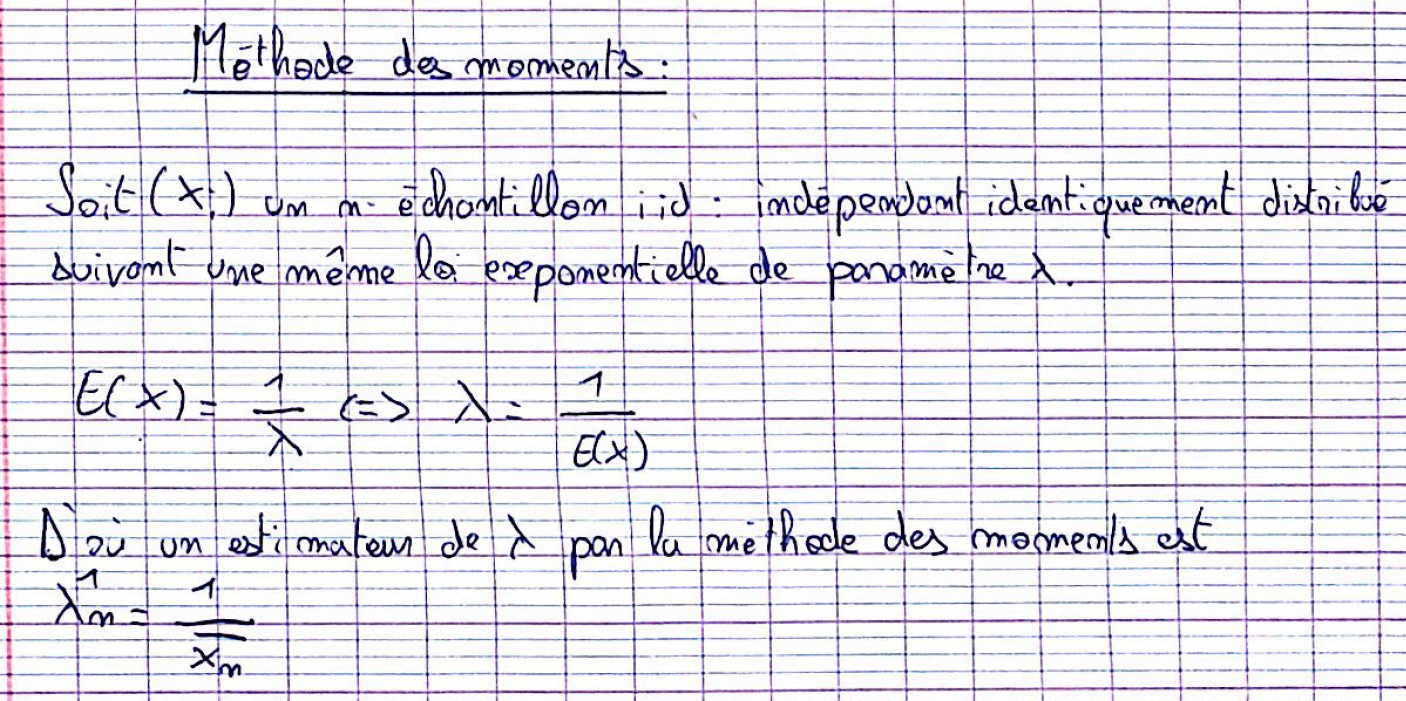
Description générée automatiquement

Nous obtenons à chaque fois une p-value inférieure à 0,05 et nous pouvons donc rejeter l’hypothèse selon laquelle elles suivent une loi normale.

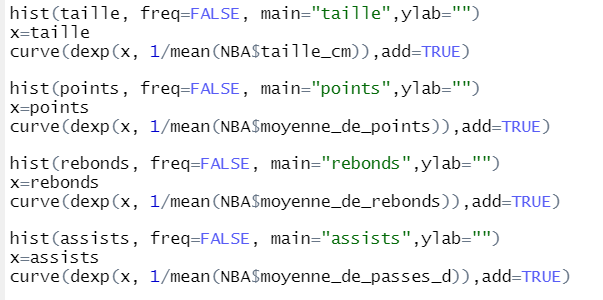
Essayons maintenant de voir si elles suivent une loi exponentielle :

Pour avoir une loi exponentielle E(λ) la plus adaptée à notre contexte il faut que nous estimions le paramètre λ de cette loi. Nous avons estimé ce paramètre par la méthode de la fonction de vraisemblance ainsi que par la méthode des moments :





Nous pouvons donc reprendre les moyennes que nous avions calculées avant pour avoir une estimation de λ pour chaque variable. Nous allons faire une analyse graphique pour savoir si nos variables semblent donc suivre une loi exponentielle. Nous avons commencé par faire la méthode de l’histogramme avec donc un histogramme et une courbe nous donnant l’allure des données pour chaque variable avec le code suivant :



Une image contenant graphique

Description générée automatiquementUne image contenant graphique

Description générée automatiquementEt nous obtenons les résultats suivants :

Une image contenant graphique

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement

Nous remarquons grâce à ces graphiques que pour l’instant les trois variables de niveau semblent suivre une loi exponentielle, mais pas la taille.

Pour avoir une deuxième vérification concernant ce résultat nous avons appliqué la méthode de la fonction de répartition avec ce code :

Une image contenant texte

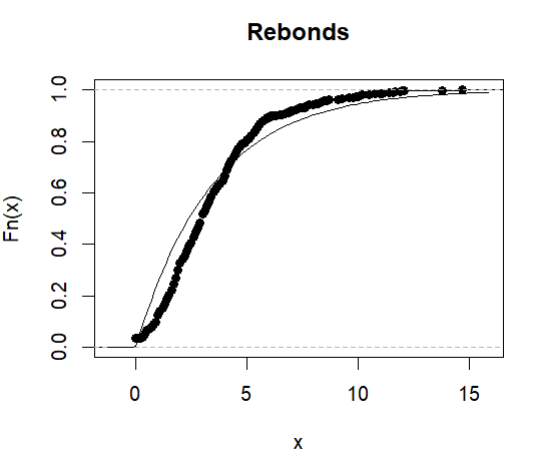
Description générée automatiquement

Ce qui nous donne un nuage de points avec la courbe associée :

Une image contenant graphique

Description générée automatiquementUne image contenant graphique

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement

Les résultats sont les mêmes que pour le test précédent, les trois variables de niveau semblent suivre une loi exponentielle, contrairement à la taille.

Les tests étant concluants pour 3 de nos variables, nous pouvons en conclure que le nombre moyen de points par matchs suit une loi exponentielle de paramètre λ=8.24, que le nombre moyen de rebonds par match suit une loi exponentielle de paramètre λ=3.446942, que le nombre moyen d’assists par match suit une loi exponentielle de paramètre λ=1.868264 et enfin que la taille ne semble pas suivre de loi connue.

Maintenant que nous avons étudié plus spécifiquement nos variables, nous allons essayer de déterminer s’il y a vraiment un gabarit parfait au basket selon nos données de la saison 2021-2022. Vu que nous avons remarqué que nos données sont écartées et qu’il n’y a pas de tendance pour l’instant, nous avons décidé de créer des nouvelles bases de données réduites, avec des contraintes concernant les points, les rebonds et les assists. Nous ne garderons que les meilleurs joueurs dans chaque domaine pour voir si dans ce cas une tendance se dessine. Pour avoir des résultats robustes nous avons d’abord créé une base de données uniquement avec les joueurs ayant joué au moins 30 matchs dans la saison car supposons qu’un joueur ait fait de très bonnes performances mais n’ait joué que quelques matchs, cela ne serait pas représentatif sur toute une saison (82 matchs).

Une image contenant texte

Description générée automatiquement

Après avoir créé cette nouvelle base de données, nous avons donc créé de nouvelles bases avec une dizaine de joueurs qui sont les meilleurs dans chacun des 3 domaines :

Une image contenant texte

Description générée automatiquement

Nous avons commencé avec le nombre moyen de points par match, avec comme joueurs tous ceux qui avaient une moyenne de points par match supérieure ou égale à 25 :

Une image contenant graphique

Description générée automatiquement

Lorsqu’on regarde le nuage de points, on ne voit pas de taille “idéale” pour ce domaine.

Nous avons répété l’opération mais cette fois en se concentrant sur le nombre moyen de rebonds par match, nous avons gardé les joueurs ayant minimum 10 rebonds par match :

Une image contenant graphique

Description générée automatiquement

Ici les résultats sont plus convaincants, on observe qu’il y a 7 joueurs sur 14 qui mesurent 211 cm et 4 joueurs sur 11 qui mesurent 209 cm. On remarque donc ici qu’une très grande partie des meilleurs joueurs en termes de rebonds suivent bien une tendance qui est celle de mesurer 210 cm (pour arrondir). Effectuons des tets pour étudier cette hypothèse :



Une image contenant texte, lettre

Description générée automatiquement

On observe que cette taille ne fait pas partie de l’intervalle de confiance à 95% de la taille donc ce sont vraiment des joueurs ayant une taille spécifique qui réussissent bien dans ce domaine et pas des joueurs lambdas. De plus, nous avons une p-value très faible ce qui nous assure que notre résultat est vraiment significatif.

Enfin, nous répétons l’opération une troisième fois mais cette fois avec le nombre moyen d’assists par match :

Une image contenant graphique

Description générée automatiquement

Nous observons un résultat semblable à celui des points par match, les observations sont plutôt

espacées, bien qu’il y en ait plusieurs entre 190 et 200 cm, ce n’est pas très précis.

Conclusion :

Nous pouvons donc conclure de ces observations sur la saison 2021-2022 qu’il n’y a pas de gabarit “parfait” pour jouer en NBA actuellement concernant les compétences techniques (points par match et assists par match), cependant s’il on veut espérer être bon dans les compétences plus physiques comme les rebonds, il faudrait mesurer 2m10 pour 110 kg environ.

# BIBLIOGRAPHIE

Notre base de données sur kaggle :

<https://www.kaggle.com/datasets/justinas/nba-players-data>

Les statistiques sur R :

<https://chesneau.users.lmno.cnrs.fr/adequation-R.pdf>

Cours de statistiques de 2 année de préparation intégrée :

<https://devinci-online.brightspace.com/d2l/le/lessons/89235/topics/304796>

<https://devinci-online.brightspace.com/d2l/le/lessons/89235/topics/304798>

<https://devinci-online.brightspace.com/d2l/le/lessons/89235/topics/310861>

Sites pour nous informer sur le basketball :

<https://fr.wikipedia.org/wiki/Basket-ball>

<https://www.parlons-basket.com/category/nba/>