MSI – TP4 : données NBA All Star Game

CLOU Noémie
FONTAINE Pierre

Groupe 9
2014-2015

Question 1 :

Le jeu de donné, classé par ordre alphabétique, présente l'ensemble des joueurs ayant participé au NBA All Star Game entre 1950 et 2009 : nom, prénom, année de participation, statistique du match, etc.

Question 2 :

Il s'agit de sélectionner l'année 2001, puis de déterminer les scores des équipes « West » et « East ». Par exemple pour l'équipe « West », on écrit le code suivant :

```
A = (which (donnees$seadon_id == 2001 & donnees$conference == "West")
pointwest = donnees$points[A]
sum(pointwest)
```

On obtient alors un score de 135 (West) à 120 (East).

Question 3 :

Pour déterminer le joueur qui a établi le record de point sur un match, on

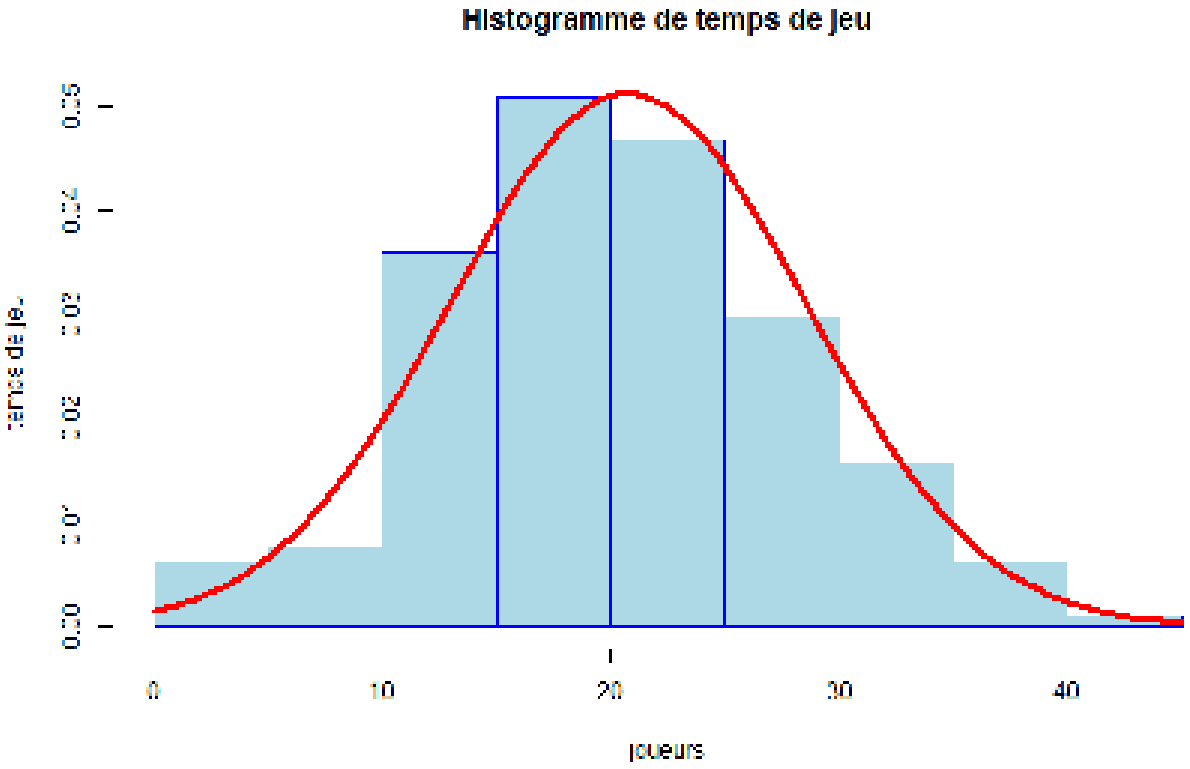
```
Q3 = which (donnees$points = max (donnees$points))
name = donnees [Q3,2]
annee=donnees[Q3,5]
```

C'est alors « chambwi01 » qui a établi le record de points en 1961.

Question 4 :

L'histogramme de densité des minutes jouées et la loi normale s'obtiennent avec le code suivant :

```
xbar = mean (donnees$minutes)
sigmabar = sd (donnees$minutes)
hist (donnees$minutes, freq = FALSE, border = "blue", col = "lightblue", xlab = "joueurs", ylab = "temps de jeu", main = "Histogramme de temps de jeu")
curve = dnorm (x,xbar,sigmabar), add = TRUE, col = "red", lwd = 4)
```

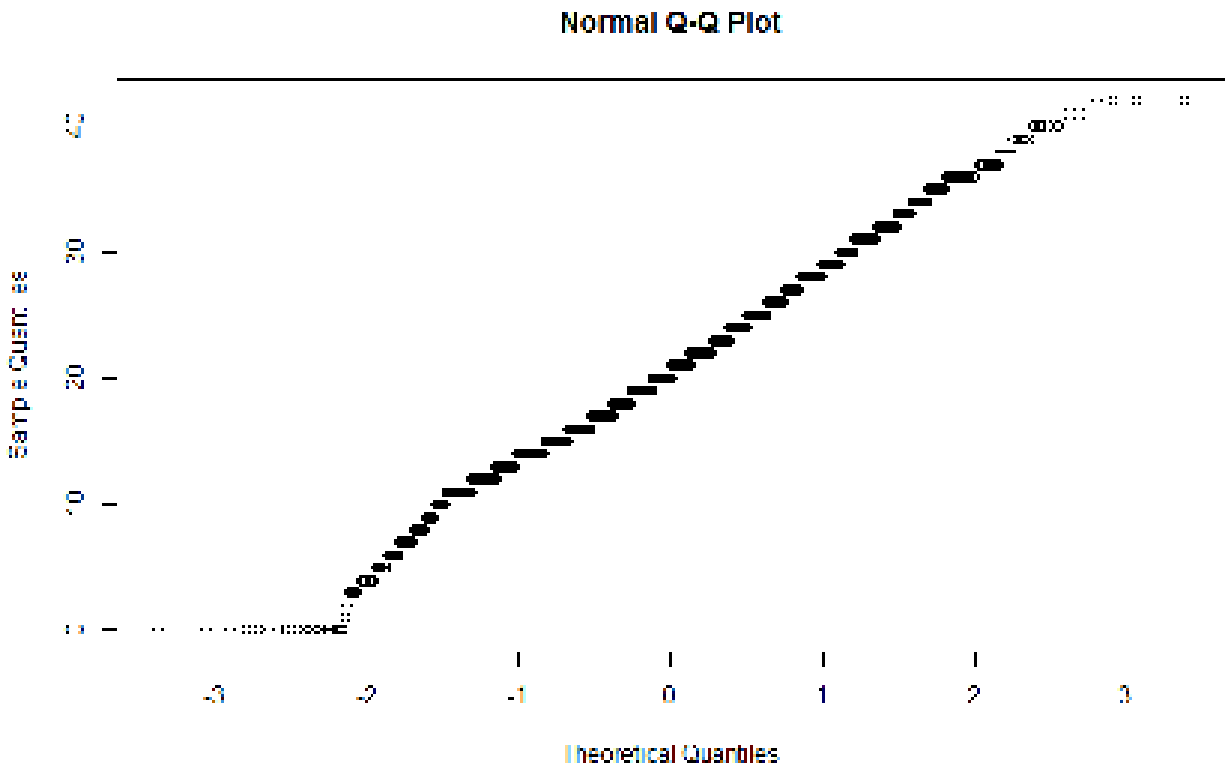


Question 5 :

On utilise le test de Shapiro : *shapiro.test (donnees$minutes)*

On obtient W = 0,9931 et pc = $6{,}23.10^{-6}$, très inférieur à $\alpha = 0{,}05$. Ainsi, l'hypothèse (H0) est rejetée, donc le nombre de minutes jouées par un jour ne suit pas une loi normale.

En utilisant la fonction *qqnorm*, le graphique des quartiles théoriques en fonction de ceux de l'échantillon n'est pas une droite pour toutes les valeurs. En effet, il n'y a pas de relation de proportionnalité pour des valeurs très faibles et très élevées, ce qui montre que le temps de jeu d'un joueur ne suit pas une loi normale.



Question 6 :

Pour répondre à cette question, nous écrivons le code suivant permettant de prendre en compte toutes les conditions fixées par l'énoncé :

Q6 = which (donnees$minutes > 15)
QQ6 = donnees$points [Q6]
t.test (QQ6, mu = 12, conf.level = 0.99)

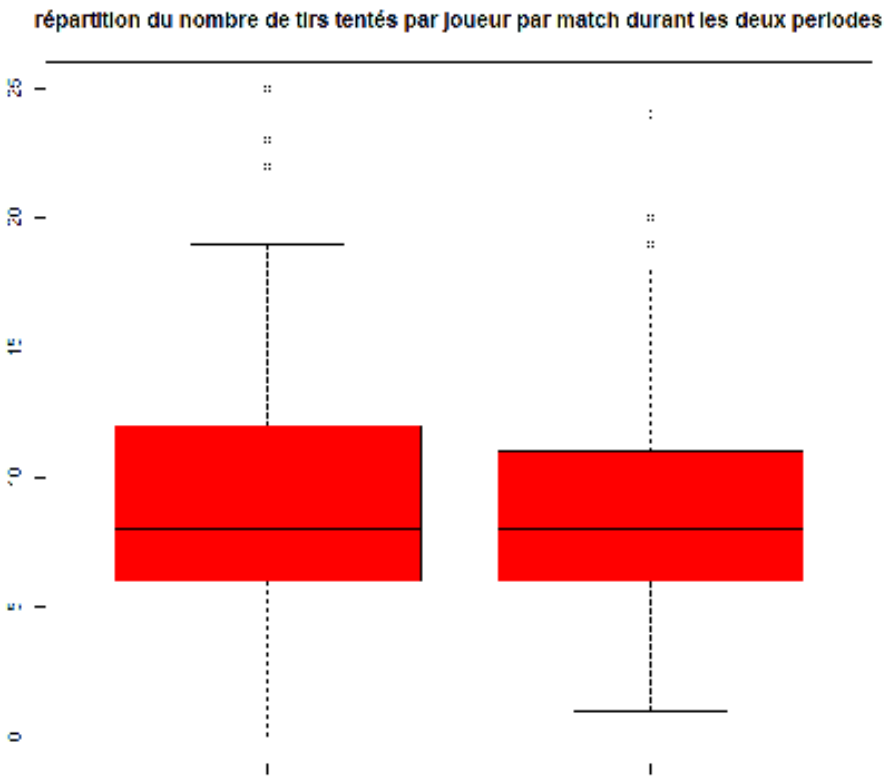
On obtient pc = 0,024 inférieur à 0,025, donc l'hypothèse (H0) n'est pas validée. On note également l'intervalle de confiance à 99 %, compris entre 11,93 et 13,04 minutes.

Question 7 :

Pour dessiner les boîtes à moustaches du nombre de tirs tentés sur les périodes 1980-1990 et 1990-2000, on écrit le code suivant :

perdiode1 = which (1980 < donnees$season_id & donnees$season_id < 1990)
perdiode2 = which (1990 < donnees$season_id & donnees$season_id < 2000)
tir1 = donnees$fg_attempted [periode1]
tir2 = donnees$fg_attempted [periode2]
boxplot (tir1, tir2, col = "red", main = "Répartition du nombre de tirs tentés par joueur par match durant les deux periodes")

On obtient alors les boîtes à moustaches suivantes :



On remarque que la moyenne reste la même entre 1980-1990 et 1980-2000, mais moins de joueurs ont tenté un nombre de paniers élevés sur la seconde période (3ème quartile plus faible et raccourcissement de la boîte à moustache).

On effectue le test de comparaison de population de Fisher-Snedecor pour la variance (var.test) et la moyenne (t.test). On remarque que le rapport des variance est égal à 1,38, ce qui confirme que les boîtes à moustaches n'ont pas les mêmes étalement et intervalle interquartile. Concernant les moyennes, elles sont très proches (9,27 pour la 1ère période et 8,82 pour la 2nde), ce qui confirme la proximité des moyennes observées sur les boîtes à moustaches.

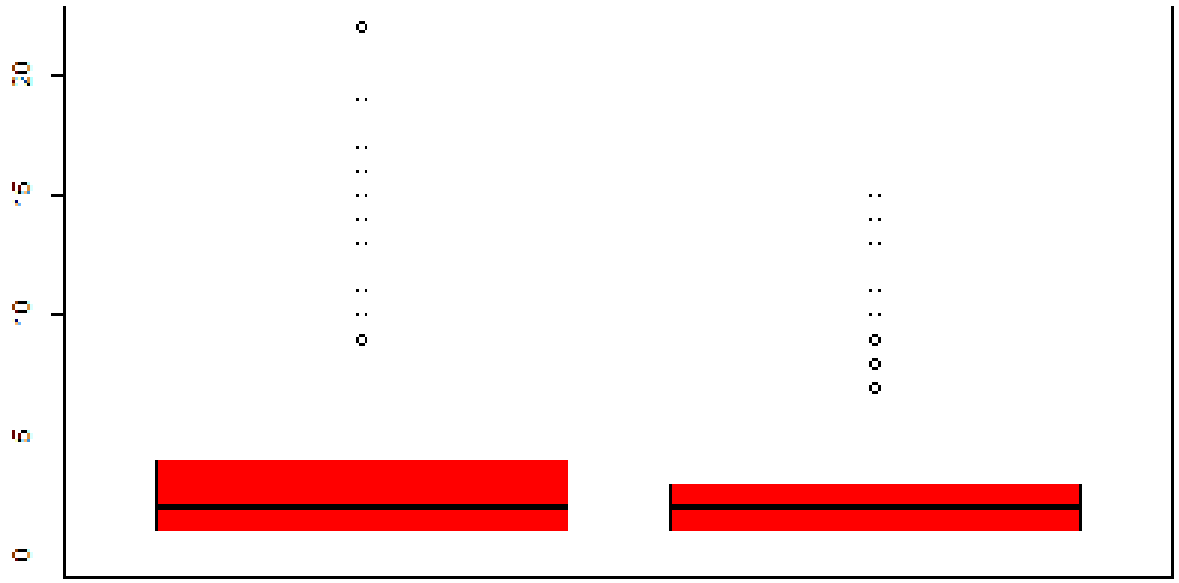
Question 8 :

De même qu'à la question précédente, en utilisant le code suivant :

awest = which (donnees$conference == "West")
aeast = which (donnees$conference == "East")
assist1 = donnees$assists [awest]
assist2 = donnees$assists [aeast]
boxplot (assist1, assist2, col = "red", main = "Répartition du nombre de passes décisives par joueur par match pour chaque conférence ")

On obtient alors les boîtes à moustaches suivantes :

répartition du nombre de passes decisives par joueur par match pour chaque conférence



En utilisant les fonctions permettant de calculer les intervalles interquartiles (*IQR*) et le test de Fisher-Snedecor (*t.test* pour la moyenne et *var.test* pour la variance), on peut tirer les conclusions suivantes :

- Les moyennes sont quasi identiques (2,66 et 2,44)
- Les intervalles interquartiles sont différents (3 pour « West » et 2 pour « East »)
- L'étalement des boîtes à moustaches est aussi plus élevé pour « West » que pour « East »

Cela confirme bien que la conférence « West » a toujours comporté plus de joueurs altruistes, c'est-à-dire des joueurs faisant le plus de passes décisives.

Question 9 :

Pour déterminer le joueur qui a participé le plus grand nombre de fois au NBA All Star Game, on utilise d'abord la fonction *unique* pour obtenir le nombre de joueurs différents (au nombre de 375). Il faudrait ensuite réussir à sommer la colonne « participation » pour chaque joueur et en sortir celui qui a la valeur la plus élevée.

Question 10 :

Calculer la puissance d'un test permet d'évaluer son efficacité, relativement à d'autres tests. Pour le rendre plus puissant, on peut soit augmenter la taille de l'échantillon, soit en diminuant l'erreur de seconde espèce.

Question 11 :

Nous utilisons désormais la fonction *power.t.test* pour la suite des questions, en utilisant les données de l'énoncé et en notant « NULL » pour la variable recherchée, comme ci-dessous :

Q11 = power.t.test (n = NULL, delta = 1, sd = 1.7, sig.level = 0.05, power = 0.8, type = c("two.sample", "one.sample", "paired"), alternative = c("two.sided", "one.sided"), strict = FALSE)

On obtient alors n = 46,35 soit un échantillon de 47 joueurs.

Question 12 :

Si on diminue le risque, en augmentant la puissance du test à 90 %, on obtient n = 61,7. L'échantillon augmente donc fortement lorsque l'on diminue le risque.

Question 13 :

En faisant de même avec n = 20, on obtient une puissance de test de 44 %.

Question 14 :

La différence de moyenne détectable à 80 % avec 20 joueurs est alors égal à 1,55.