TP 3 : Estimation et Test

Pour ce TP nous allons reprendre les données sur la position des palindromes dans l'ADN du virus CMV. Ces données nous serviront de support afin de faire nos premiers ajustements de loi de probabilités et tests d'hypothèses.

1 Inspection des données

Commençons par importer les données qui sont contenues dans le fichier `hcmv.data` et qui contient la position des palindromes — pour la première paire de base. Une fois les données importées et attachées, lancez la commande suivante.

```
plot(location %% 50000, location %/% 50000, xlab = "Position dans l'ADN",
     ylab = "", pch = 3, yaxt = "n")
axis(2, at = 0:4, labels = seq(0, 200000, by = 50000), las = 2)
```

Que représente cette figure ?

Analysez la distribution de la distance entre deux palindromes successifs — la fonction `diff` pourra vous être utile.

On peut également s'intéresser à la distribution du nombre de palindromes dans des segments d'ADN disjoints de longueur 4000 paires de bases. Ceci peut-être fait par

```
segments <- seq(1, 229354, by = 4000)
effectifs <- table(cut(location, breaks = segments, labels = FALSE))
effectifs
```

Que donne cette sortie ? Comparer avec les données du cours. Représentez graphiquement ces effectifs — mais prudence ! Commentez.

2 Premières analyses

D'après le cours nous savons que le processus de Poisson (homogène) suppose que la position est uniformément répartie. Sur quelques tentatives, représentez la position des palindromes avec des positions de palindromes uniformément réparties — avec bien entendu le même nombre total de palindromes. Rappel : l'ADN du CMV contient 229 354 paires de base. Par exemple, on pourra utilisez le code suivant

```
n.palindromes <- length(location)

par(mfrow = c(1,4))
plot(location %% 50000, location %/% 50000, xlab = "Position dans l'ADN",
     ylab = "", pch = 3, yaxt = "n", main = "Observations")
axis(2, at = 0:4, labels = seq(0, 200000, by = 50000), las = 2)

for (i in 1:3){
  unif.loc <- sort(sample.int(229354, n.palindromes))
  plot(unif.loc %% 50000, unif.loc %/% 50000, xlab = "Position dans l'ADN",
       ylab = "", pch = 3, yaxt = "n", main = "Simulations")
  axis(2, at = 0:4, labels = seq(0, 200000, by = 50000), las = 2)
}
```

Visuellement que pouvez vous dire ?

Ici nous avons simulé que 3 échantillons mais la puissance de l'ordinateur nous permet de le faire un grand nombre de fois. Il pourrait être intéressant de comparer la distribution de la distance entre deux palindromes observées et à celle de la distance moyenne sur nos échantillons simulés. Le code suivant permet une telle analyse.

```
distance <- diff(location)
n.sim <- 500
distance.sim <- matrix(NA, n.sim, length(distance))

for (i in 1:n.sim){
  unif.loc <- sort(sample.int(229354, n.palindromes))
  distance.sim[i,] <- sort(diff(unif.loc))
}

qqplot(distance, colMeans(distance.sim), xlab = "Distances observees",
       ylab = "Distances simulees")
abline(0, 1)
```

Que pouvez vous dire à partir de ce graphique ?

3 Test d'adéquation

L'objectif de cette partie est de vérifier si le processus de Poisson (homogène) permet de bien représenter nos données. Le processus de Poisson suppose plusieurs choses. Le nombre de palindromes pour une région donnée suit une loi de Poisson de paramètre noté λ. De plus le nombre de palindromes pour deux segements d'ADN disjoints sont indépendants. Enfin le nombre de palindromes étant fixé, la position de ces palindromes est alors répartie uniformément.

A l'aide du cours, comment estimeriez vous λ, enregistrez cette valeur dans un objet `lambda.hat`.

Que proposeriez vous de faire pour tester si les données sont bien représentées par un processus de Poisson ?

Nous allons implémenter ce qui doit être votre idée je l'espère. Les effectifs observés ainsi que le probabilités d'appartenir à chacune des classes utilisées en cours peuvent être obtenus par

```
classes <- c(-Inf, 2, 3, 4, 5, 6, 7, 8, Inf)
eff.obs <- table(cut(effectifs, breaks = classes))
prob.theo <- diff(ppois(classes, lambda.hat))
```

Comparez ces résultats avec ceux du cours. Que renvoie la commande

```
chisq.test(eff.obs, p = prob.theo)
```

Dans la sortie précédente, que signifie la sortie `df=7`. Est-ce en accord avec la théorie vue en cours ? Pourquoi ? Que nous indique le message d'avis ? Comment feriez vous pour le supprimer ? Comment calculeriez vous la vraie p-valeur ? Enfin quelles conclusions pouvez vous tirer de ce test d'hypothèse.

4 Nombre de palindromes anormalement élevés

Certains chercheurs supposent que l'origine de réplication du virus est codée par une concentration anormalement élevée de palindromes. Afin d'identifier ces zones dans l'ADN, nous proposons le graphique suivant.

```
debut <- seq(1, 229354, by = 500)
n.segment <- length(debut)

nb.palindrome <- rep(NA, n.segment)
for (i in 1:n.segment)
  nb.palindrome[i] <- sum((location >= debut[i]) & (location <= (debut[i] + 999)))

plot(seq(500, 229500, by = 500), nb.palindrome, type = "l", xlab = "Position",
     ylab = "Nombre de palindromes")
```

Essayez de comprendre ce que représente ce graphique. Utilisez la fonction `max` afin de déterminer le nombre maximum de palindromes dans un segment ayant 1000 paires de base.

Afin de déterminer si ce nombre maximal est anormalement élevé, on propose de simuler un grand nombre de position selon le processus de Poisson. Faites tourner le code suivant et interprétez.

```
n.sim <- 1000

max.sim <- matrix(NA, n.sim)
for (i in 1:n.sim){
  loc.sim <- sort(sample.int(229354, n.palindromes))

  nb.palindrome.sim <- rep(NA, n.segment)
  for (j in 1:n.segment)
    nb.palindrome.sim[j] <- sum((loc.sim >= debut[j]) & (loc.sim <= (debut[j] + 999)))

  max.sim[i] <- max(nb.palindrome.sim)
}

hist(max.sim, freq = FALSE, xlab = "Nombre maximal de palindromes (simulations)",
     main = "")
```