TP5 - éléments de correction

Exercice 1

On commence par copier les données :

```
clients <- 0:6
magasins <- c(37,46,39,19,5,3,1)
```

Pour estimer l'espérance et la variance, on peut reproduire des données brutes correspondant à la situation décrite dans le tableau.

```
donnees <- rep(clients, magasins)
mu <- mean(donnees)
sig2 <- var(donnees)
```

Il s'agit de comparer les effectifs observés et les effectifs théoriques. Autrement dit, on compare les deux distributions de fréquences en faisant intervenir l'effectif total, qui influencera la significativité des différences observées.

On se souvient qu'une variable aléatoire X suivant une loi de Poisson de paramètre λ vérifie :

$$\lambda = \mathbb{E}(X) = \mathrm{Var}(X)$$

Si la loi sous-jacente est bien une loi de Poisson, on doit déjà avoir une moyenne et une variance proches l'une de l'autre.

```
mu
```

```
## [1] 1.48
```

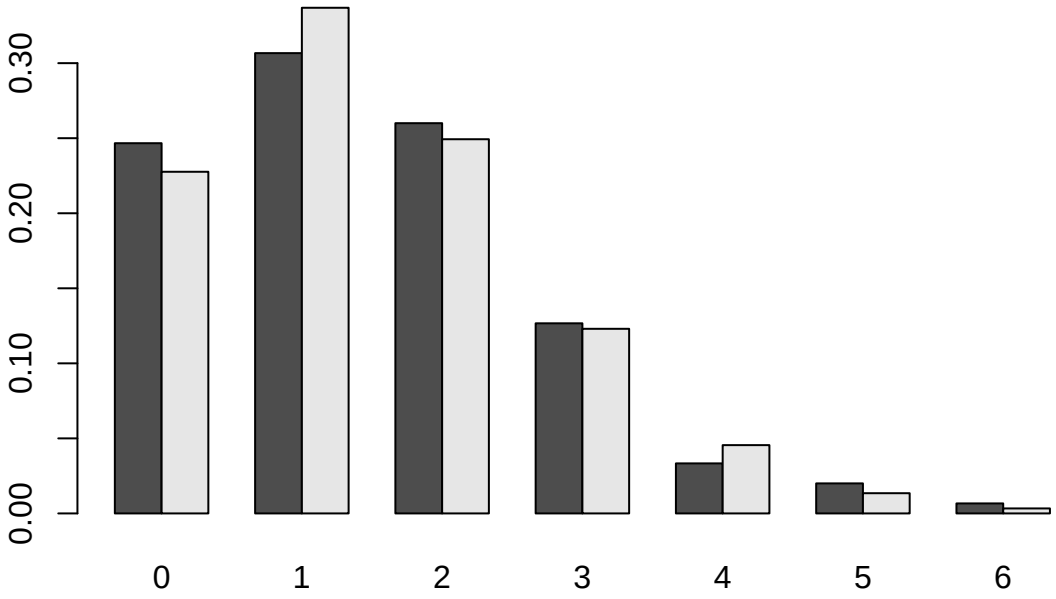
```
sig2
```

```
## [1] 1.580134
```

C'est le cas : c'est bon signe.

On peut représenter la situation par un double histogramme à barres accolées :

```
freq.obs <- magasins/sum(magasins)
freq.theo <- dpois(0:6, lambda = mu)
barplot(rbind(freq.obs,
              freq.theo),
        names.arg=clients,
        beside=T)
```



On voit que les deux distributions sont proches. Il s'agit maintenant de réaliser un test du χ^2 pour déterminer si ces faibles différences sont significatives ou non. Pour ce faire, on doit comparer le tableau des effectifs associés à cette situation : on compare effectifs observés et effectifs théoriques.

```
tab <- rbind(freq.obs, freq.theo)*150
```

On multiplie par 150 car c'est l'effectif total. On passe de fréquences à effectifs en multipliant par l'effectif total.

Réalisons maintenant un test du χ^2, dont les hypothèses peuvent s'écrire :

$$H_0 \ : \ X \sim \mathcal{P}(\mu)$$
$$H_1 \ : \ X \not\sim \mathcal{P}(\mu)$$

Plus rigoureusement, on teste en fait l'indépendance des facteurs "nombre de clients" et "effectifs observés/effectifs théoriques". S'ils sont indépendants, on a bien adéquation entre les deux distributions.

```
chisq.test(tab)
```

```
## Warning in chisq.test(tab): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 1.0103, df = 6, p-value = 0.9852
```

Au seuil de significativité $\alpha = 0.05$, ce test n'est pas significatif. On conserve l'hypothèse nulle H_0 et on décide que la loi suivie par ces données est bien une loi de Poisson $\mathcal{P}(1.48)$ (on a $\mu = 1.48$).

On pourrait également tester l'adéquation entre une $\mathcal{P}(\text{Var}(X))$ et nos données, puisque *dans ce cas* la variance comme l'espérance sont des estimateurs du paramètre λ.

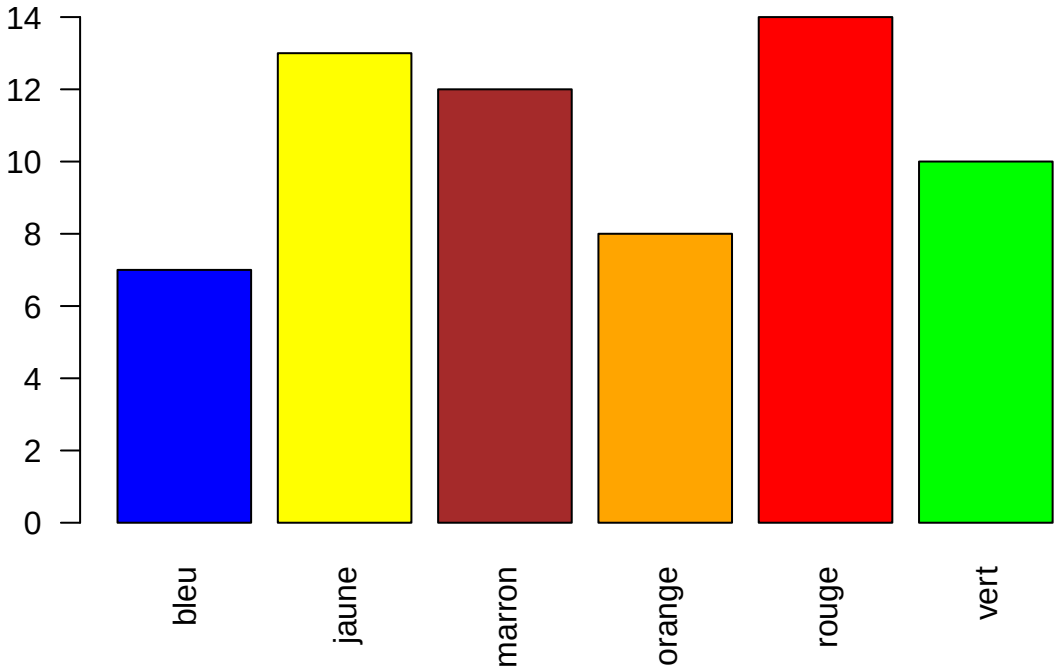
Exercice 2

On charge les données :

```
bonbons <- read.csv("~/Desktop/bonbons.csv")
```

Une représentation graphique adaptée est la suivante :

```
barplot(table(bonbons$couleur),
        las=2,
        col=c("blue","yellow","brown","orange","red","green"))
```



Les hypothèses sont semblables à celles de l'exercice précédent :

$$H_0 \text{ : les facteurs "observé/théorique" et "couleur" sont indépendants}$$
$$H_1 \text{ : les facteurs "observé/théorique" et "couleur" ne sont pas indépendants}$$

Pour réaliser ce test, on peut construire le tableau associé :

```
tab <- rbind(table(bonbons$couleur),
             c(0.2,0.1,0.2,0.1,0.3,0.1)*64)
```

Comme dans l'exercice précédent, on multiplie par 64 car c'est l'effectif total. On souhaite comparer les effectifs théoriques et observés.

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 5.7526, df = 5, p-value = 0.331
```

Le test n'est pas significatif. L'hypothèse selon laquelle la distribution donnée dans l'énoncé est celle suivie par les quantités de bonbons de différentes couleurs est conservée.

Testons l'hypothèse d'uniformité. Plus précisément :

$$H_0 \ : \text{ les proportions de chaque couleur sont égales}$$

$$H_1 \ : \text{ il existe au moins une couleur de proportion différente des autres}$$

```
tabb <- rbind(table(bonbons$couleur),
              rep(1/6,6)*64)
chisq.test(tabb)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabb
## X-squared = 1.9224, df = 5, p-value = 0.8598
```

Le test n'est pas significatif. L'hypothèse d'uniformité est conservée. On observe d'ailleurs que la $p-$valeur est plus élevée ici. On peut dire que l'hypothèse de normalité est meilleure que l'hypothèse proposée par notre ami.

Exercice 3

Chargeons le jeu de données.

```
salaries <- read.csv("~/Desktop/salaries.csv", row.names=1)
```

On va se concentrer sur les variables qualitatives (`rank`, `discipline`, `sex`) et la variable quantitative `salary`.

Révisons nos tests paramétriques

On souhaite savoir si, dans cette université, il existe une différence de salaire significative entre les hommes et les femmes. Pour savoir quel test utiliser (Student ou Welch), il faut d'abord déterminer si les variances des deux groupes sont égales ou non. Choisissons un seuil de significativité classique : $\alpha = 0,05$.

```
var.test(salaries$salary ~ salaries$sex)
```

```
##
##  F test to compare two variances
##
## data:  salaries$salary by salaries$sex
## F = 0.72702, num df = 38, denom df = 357, p-value = 0.2309
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4718089 1.2292716
## sample estimates:
## ratio of variances
##          0.7270166
```

La $p-$valeur est d'environ $0,23 > 0,05$. Le test n'est donc pas significatif, on décide de considérer les variances égales.

Réalisons le test de Student.

```
t.test(salaries$salary ~ salaries$sex,
         var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  salaries$salary by salaries$sex
## t = -2.7817, df = 395, p-value = 0.005667
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -24044.910  -4131.107
## sample estimates:
## mean in group Female   mean in group Male
##             101002.4             115090.4
```

La $p-$valeur étant inférieure au seuil de significativité $\alpha = 0.05$, le test est significatif. On a détecté une différence significative entre les salaires des femmes et ceux des hommes dans cette université. Une étude approfondie permettrait d'identifier les facteurs influençant cette différence, mais ces méthodes ne sont pas au programme cette année.

On souhaite tester l'indépendance de deux variables qualitatives. Par exemple, testons l'indépendance des variables `rank` et `discipline`.

Il nous est nécessaire de construire le tableau de contingence associé :

```
tab <- table(salaries$rank, salaries$discipline)
tab
```

```
##
##                A   B
##   AssocProf   26  38
##   AsstProf    24  43
##   Prof       131 135
```

Testons maintenant l'indépendance :

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 4.6487, df = 2, p-value = 0.09785
```

La $p-$valeur est d'environ $0,098$. Le test est donc significatif au seuil $\alpha = 0.1$, mais non-significatif au seuil $\alpha = 0.05$. Dans le premier cas, on décide que les variables sont dépendantes (le grade d'un chercheur et son domaine de recherche sont liés), alors que dans le second on décide que ces variables sont indépendantes.

Vous pouvez vous exercer sur les autres couples de variables qualitatives.