

Devoir « Compléments de statistique » du 12 Janvier 2011

Utilisation du logiciel

Description des données : Les données sont issues du « Mozilla Open Data Visualization Competition »¹, version 2010 et, plus précisément, de la plateforme de recherche « Test Pilot » qui recueille des données utilisateur sur Firefox. Les données utilisées sont les données « A Week in the Life of a Browser - Version 2: Aggregated Data Samples » qui sont disponibles à l'adresse : <https://testpilot.mozillalabs.com/testcases/a-week-life-2/aggregated-data.html>. À partir des données originelles, nous avons conservé uniquement les données relatives aux réponses de l'enquête menée sur l'utilisation de Firefox (et pas les données événementielles ni les données utilisateur) et nous nous sommes restreints aux données relatives à des utilisateurs du système d'exploitation Mac OS X. Le fichier de données final, que vous allez charger et sur lequel vous allez travailler, contient 397 individus (des utilisateurs Firefox sous système d'exploitation sous Mac OS X) et 14 variables (les réponses aux 14 questions de l'enquête).

Instructions pour commencer : Les fichiers à utiliser pour le devoir se situent dans le répertoire utilisateur /home/stid/DevoirNVV-NOM² où NOM est votre nom. Vous y trouverez le fichier Devoir-NOM.odt, qui est le fichier OpenOffice.org (traitement de texte) que vous devez remplir et rendre en fin de devoir, et le fichier FirefoxSurvey.RData qui est le fichier contenant les données. Pour commencer :

- Ouvrez le fichier Devoir-NOM.odt ;
- Lancer R et le package Rcmdr ;
- Définir le répertoire courant (c'est donc /home/stid/DevoirNVV-NOM) ;
- Ouvrir le fichier de données : il contient le tableau de données, nommé surveymac, dans lequel sont situées les données décrites ci-dessus. Les colonnes de ce tableau sont nommées q1, q2 ... q14

N'oubliez pas d'enregistrer régulièrement votre travail !

Orientation générale du devoir : Le devoir portera sur la question générale suivante : « Existe-t-il des comportements différenciés selon le sexe concernant l'utilisation d'internet ? » En particulier, on s'intéressera au fait d'utiliser plusieurs navigateurs et à la durée passée sur internet. Les questions de l'enquête étudiées seront donc :

- la question q2 : « Utilisez-vous plus d'un navigateur dans votre vie quotidienne ? » (réponses possibles : « Non. J'utilise seulement Firefox », codée 0 et « Oui, j'utilise d'autres navigateurs que Firefox », codée 1) ;
- la question q5 : « Quel est votre sexe ? » (réponses possibles : « Homme », codée 0 et « Femme », codée 1) ;
- la question q7 : « Combien de temps passez-vous sur internet par jour ? » (réponses possibles : « Moins d'une heure », codée 0, « 1-2 heures », codée 1, « 2-4 heures », codée 2, « 4-6 heures », codée 3, « 6-8 heures », codée 4, « 8-10 heures », codée 5 et « plus de 10 heures », codée 6).

1 <http://design-challenge.mozillalabs.com/open-data/OpenDataCompetition.php>

2 Les étudiants utilisant leur ordinateur personnel trouveront le fichier DevoirNVV-NOM sur le bureau de leur machine.

Table des matières

1Analyse univariée de la variable « Sexe » (Question q5).....	2
2Analyse bivariée des variables « Utilisation d'autres navigateurs » (Question q2) et « Sexe ».....	3
3Analyse univariée de la variable « Temps passé sur internet » (Question q7).....	5
4Analyse bivariée des variables « Durée journalière passée sur internet » et « Sexe ».....	5

1 Analyse univariée de la variable « Sexe » (Question q5)

1. Recodez la variable « q5 » (dans une nouvelle variable que vous appellerez « gender ») de manière à ce que les modalités 0 et 1 soient remplacées, respectivement, par « Homme » et « Femme ». La variable « gender » ainsi générée doit être un facteur.

Code généré :

```
surveymac$gender <-  
  recode(surveymac$q5,'0="Homme";1="Femme";',as.factor.result=TRUE)
```

2. **Si vous n'avez pas réussi à effectuer la question précédente**, tapez :

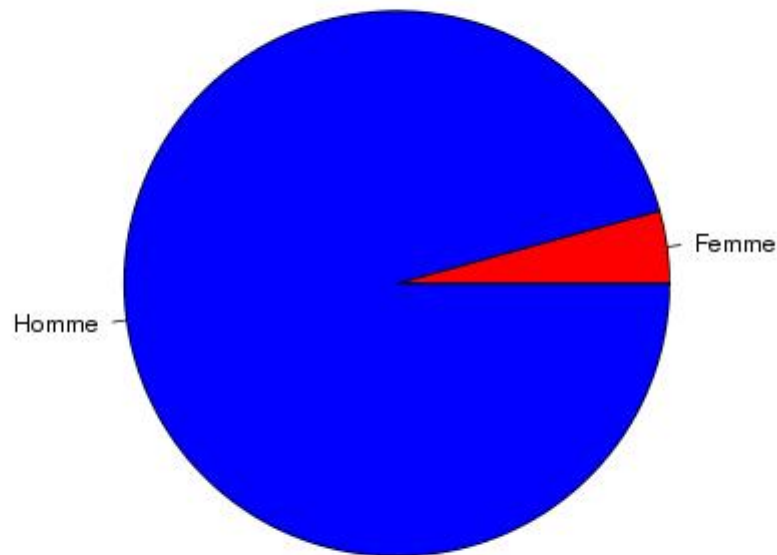
```
surveymac$q5 <- as.factor(surveymac$q5)
```

pour pouvoir continuer. Ceci n'est **pas** la réponse à la question précédente mais vous permettra d'utiliser la variable q5, non recodée, pour faire la suite.

Insérez ci-dessous le diagramme circulaire de la distribution des sexes dans l'échantillon. Commentez cette distribution.

Diagramme circulaire :

Distribution du sexe des sondés



Commentaires : La très grande majorité des sondés de l'étude (utilisateurs de Firefox sous Mac OS X) sont des hommes.

Code généré :

```
pie(table(surveymac$gender),labels=levels(surveymac$gender),main="Distribution du sexe des sondés",col=c("red","blue"))
```

2 Analyse bivariée des variables « Utilisation d'autres navigateurs » (Question q2) et « Sexe »

1. Recodez la variable « q2 » (dans une nouvelle variable que vous appellerez « other.browsers ») de manière à ce que les modalités 0 et 1 soient remplacées, respectivement, par « Non » et « Oui ». La variable « other.browsers » ainsi générée doit être un facteur.

Code généré :

```
surveymac$other.browsers <-  
  recode(surveymac$q2,'0="Non"; 1="Oui";',as.factor.result=TRUE)
```

2. **Si vous n'avez pas réussi à effectuer la question précédente**, tapez :

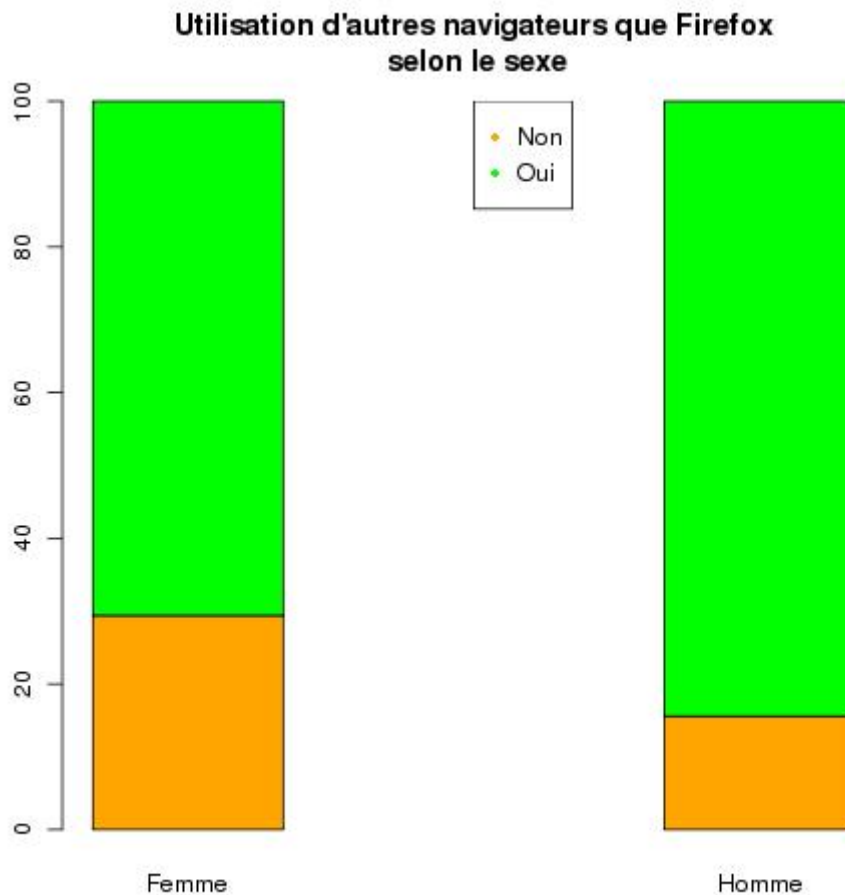
```
surveymac$q2 <- as.factor(surveymac$q2)
```

pour pouvoir continuer. Ceci n'est **pas** la réponse à la question précédente mais vous permettra d'utiliser la variable q2, non recodée, pour faire la suite.

Représentez la distribution de la variable « Utilisation d'un autre navigateur que Firefox » conditionnellement à la variable « Sexe ».

Remarque : Si, lors de la réalisation de cette question, vous avez le message d'erreur « impossible de trouver la fonction "abind" », tapez library(abind) et recommencez.

Diagramme :



Code généré :

```
tableC <- table(surveymac$gender,surveymac$other.browsers)
tableC.cond <- rowPercents(tableC)
tableC.cond <- tableC.cond[,1:2]
barplot(t(tableC.cond),col=c("orange","green"),space=2,main="Utilisation d'autres navigateurs que Firefox\n selon le sexe")
legend(4,100,legend=c("Non","Oui"),pch=18,col=c("orange","green"))
```

Qui, des hommes ou des femmes, a le plus tendance à utiliser plusieurs navigateurs ?

Réponse : La proportion d'hommes utilisant d'autres navigateurs que Firefox est plus grande que la proportion de femmes utilisant d'autres navigateurs que Firefox.

3 Analyse univariée de la variable « Temps passé sur internet » (Question q7)

On recode la variable q7 (« Durée journalière passée sur internet ») de manière à la rendre à nouveau quantitative (pour l'instant elle est regroupée en classes) en remplaçant chaque classe par son centre ; la nouvelle variable ainsi créée sera nommée « time ». Pour cela, exécutez la commande :

```
surveymac$time <-
recode(surveymac$q7,'0=0.5;1=1.5;2=3;3=5;4=7;5=9;6=12;',as.factor.result=FALSE)
```

1. Calculez la moyenne, la médiane et l'écart type de la variable « Durée journalière passée sur internet ».

Moyenne, médiane et écart type :

La durée journalière moyenne passée sur internet est 6,03 heures avec un écart type de 2,98 heures. La médiane de la durée journalière moyenne passée sur internet est 5 heures.

Code généré :

```
mean(surveymac$time)
median(surveymac$time)
sd(surveymac$time)
```

4 Analyse bivariée des variables « Durée journalière passée sur internet » et « Sexe »

1. Déterminez les moyennes et écarts types de la « Durée journalière passée sur internet » conditionnellement au « Sexe ». Commentez les différences entre hommes et femmes.

Réponse :

La durée journalière moyenne passée par les femmes sur internet est 5,85 heures (avec un écart type de 3,26) et la durée journalière moyenne passée par les hommes sur internet est 6,04 heures (avec un écart type de 2,97).

Commentaires :

Les femmes passent en moyenne moins de temps que les hommes sur internet mais il y a une plus grande dispersion dans la distribution du temps passé sur internet par les femmes que dans la distribution du temps passé sur internet par les hommes.

Code généré :

```
by(surveymac$time,surveymac$gender,mean)  
by(surveymac$time,surveymac$gender,sd)
```

2. Calculez $\eta(X|Y)$ pour X la variable « Durée journalière passée sur internet » et Y la variable « Sexe ». Commentez sa valeur.

Réponse :

$$\eta(X|Y) = 0,013$$

Commentaires :

$\eta(X|Y)$ a une faible valeur : le sexe influence peu le temps moyen passé sur internet.

Code généré :

```
AnovaModel.1 <- aov(time ~ gender, data=surveymac)  
Tableau <- summary(AnovaModel.1)[[1]]  
sqrt(Tableau[1,"Sum Sq"]/sum(Tableau["Sum Sq"]))
```