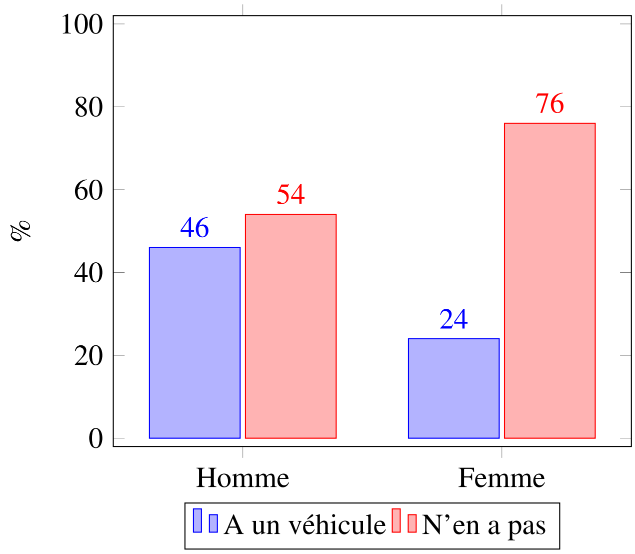
CORRIGÉ

Exercice 2 : *Croisement de deux variables qualitatives*

FIGURE 1 – PROPORTION D'ÉTUDIANTS D'AES ET DE SEG POSSÉDANT UN VÉHICULE SELON LEUR GENRE.



Source : Enquête 2015.

TABLE 1 – GENRE ET POSSESSION D'UN VÉHICULE POUR 34 ÉTUDIANTS D'AES ET DE SEG.

Indiv.	*Genre*	*Véhicule*	*Indiv.*	*Genre*	*Véhicule*
1	H	V	18	H	NV
2	H	V	19	H	V
3	F	NV	20	F	NV
4	F	NV	21	F	V
5	F	NV	22	F	NV
6	H	NV	23	F	V
7	F	NV	24	F	NV
8	F	NV	25	H	V
9	F	NV	26	F	NV
10	H	V	27	F	NV
11	H	NV	28	H	V
12	F	V	29	F	NV
13	H	NV	30	F	NV
14	F	NV	31	F	NV
15	F	NV	32	H	V
16	H	NV	33	F	NV
17	H	NV	34	F	V

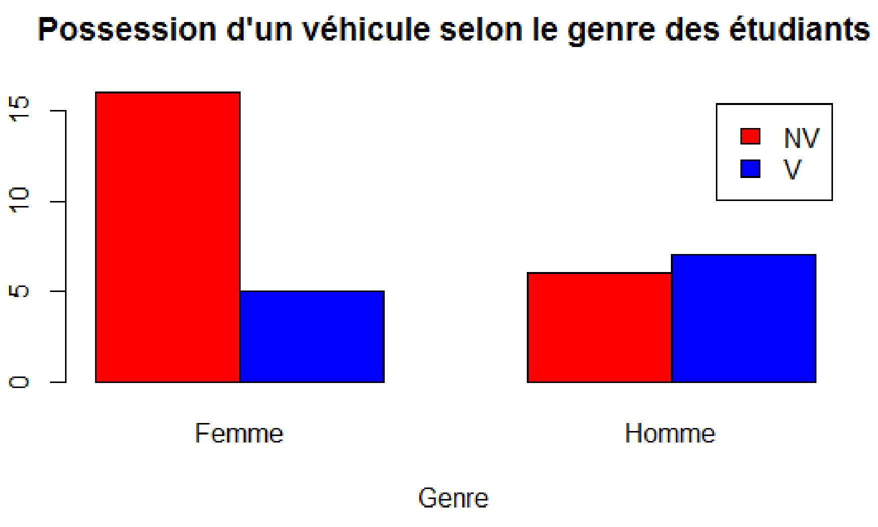
Note : Nous avons noté "V" lorsque l'étudiant possédait un véhicule, et "NV" dans le cas contraire.
Source : Enquête 2015.

1. **Saisir les données du tableau 1 sous R sous forme d'un data frame.**
```
Genre <- c("Homme", "Homme", "Femme", "Femme", "Femme",
"Homme", "Femme", "Femme", "Femme", "Homme", "Homme",
"Femme", "Homme", "Femme", "Femme", "Homme", "Homme",
"Homme", "Homme", "Femme", "Femme", "Femme", "Femme",
"Femme", "Homme", "Femme", "Femme", "Homme", "Femme",
"Femme", "Femme", "Homme", "Femme", "Femme")
Vehicule <- c("V", "V", "NV", "NV", "NV", "NV", "V",
"NV", "NV", "V", "NV", "V", "NV", "NV", "NV", "NV",
"NV", "NV", "V", "NV", "V", "NV", "V", "NV", "V", "NV",
"NV", "V", "NV", "NV", "NV", "V", "NV", "V")

Enquete <- data.frame(Genre = Genre, Vehicule = Vehicule)
```
2. **Construire le tableau de contingence croisant les deux variables.**
```
tablecroisee <- table(Enquete$Vehicule, Enquete$Genre)
tablecroisee

          Femme Homme
    NV     16     6
    V       5     7
```
3. **Reproduire la figure 1.**
```
par(mfrow=c(1,1))
barplot(tablecroisee,
main="Possession d'un véhicule selon le genre des étudiants",
xlab="Genre", col=c("red","blue"),
legend = rownames(tablecroisee), beside=TRUE)
```

FIGURE 2 – NOMBRE D'ÉTUDIANTS D'AES ET DE SEG POSSÉDANT UN VÉHICULE SELON LEUR GENRE.



Source : `Enquête 2015`.

4. **Déterminer la valeur du χ^2.**
```
resultat <- chisq.test(Enquete$Genre, Enquete$Vehicule)
resultat
```
On se souvient que :

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - \hat{n_{ij}})^2}{\hat{n}_{ij}}$$

$$\chi^2 = 1,99$$

$\chi^2 \neq 0$ donc on s'éloigne de la situation d'indépendance entre le genre des étudiants et la possession d'une voiture. Toutefois, cela manque encore de rigueur.

Bonus **Démontrer le lien de dépendance entre la possession d'un véhicule et le genre des étudiants (au seuil de 5%).** Table

```
resultat <- chisq.test(Enquete$Genre, Enquete$Vehicule)
resultat
```

En dépassant le cadre strict des statistiques descriptives, nous pouvons tester le lien de dépendance entre deux variables à l'aide du test du χ^2. Quatre étapes sont nécessaires :

(a) Formuler les hypothèses du test du χ^2 :
On pose :
- H_0 : les variables X et Y sont indépendantes.
- H_1 : les variables ne sont pas indépendantes.

(b) Détermination des degrés de liberté (ddl) :
- On détermine p le nombre de lignes et q le nombre de colonnes, dans la mesure où : la distance du χ^2 suit la loi du χ^2 à $(p-1)(q-1)ddl$.

(c) On définit le seuil (la marge d'erreur) : ici 5%.
- Cela dépend de la précision attendue par l'analyste lors du test, par exemple l'influence d'un médicament ne sera pas évaluée avec la même précision que la réussite des étudiants aux examens.
- Dans notre cas, cela signifie que nous avons 5% de chance de rejeter H_0 à tord.
- Il ne faut pas sous-estimer les probabilités faibles !

(d) Enfin, on va chercher dans la table du chi^2 afin de repérer la valeur critique au croisement de la ligne ddl et du seuil d'erreur, on va donc chercher $K_{95\%} = F(0,05;1)$

(e) On compare enfin cette statistique à la valeur du chi^2 trouvée précédemment.
- Si la valeur du χ^2 est inférieure à la valeur critique (issue de la table), cela conduit à accepter H_0 et à conclure à l'indépendance entre les deux variables.
- Inversement si la valeur est supérieure.
- Une autre façon est d'utiliser la P-value apparaissant dans la console sur R. Au seuil de 5%, si la valeur de la P-value est inférieure à 5% alors on se situe dans le premier cas.
- Inversement si la P-value est supérieure à 5%.

Dans notre cas :

$$\chi^2 = 1,99$$

Dans la console, on observe :

$$p - value = 0,158$$

On doit donc rejeter l'hypothèse H_0 et conclure à la dépendance entre les deux variables au seuil de 5%.

Corrigé [en ligne le 05/10/2016] [002A]