TP n° 3 : Distributions à deux variables qualitatives

L'objectif de ce TP est d'étudier le relation entre deux variables qualitatives. Nous allons pour cela effectuer un test d'indépendance de Chi-deux. Puis pour quantifier cette relation, nous utiliserons le coefficient de Cramer.

Le fichier "diplome_sexe.csv" recense le sexe et le niveau de diplôme obtenu d'un échantillon aléatoire de 1367 diplômés d'université.

1. Télécharger puis importer ce fichier.
2. Etudier les lignes de commande suivantes:

```
dim(data)
head(data)
diplome=data$Diplome
sexe=data$Sexe
levels(diplome)
table(diplome)
length(which(diplome=="Licence"))
```

3. Avec la fonction *table*, croiser les deux variables *diplome* et *sexe* dans un tableau de contingence qui recense les effectifs pour chaque combinaison de valeurs des variables.
4. Utiliser la fonction *chisq.test* pour conclure quant à la dépendance entre le sexe et le niveau de diplôme obtenu.

Afin d'évaluer le degré de relation entre les deux variables qualitatives, on peut utiliser l'indice de Cramer qui varie entre 0 et 1. Si le coefficient est proche de 0, les variables ne sont pas lieés (indépendantes). Si le coefficient est proche de 1, les variables sont liées (dépendantes).

$$V = \sqrt{\frac{\chi^2}{n \times \min(p-1, q-1)}}$$

5. Compléter la fonction ci-dessous qui partant de deux variables qualitatives (factor), retourne la valeur du coefficient de Cramer.

```
cramer <- function(x, y) {
res <- chisq.test(x, y)
chi2 <- as.numeric(res$statistic)
n <- ............
p <- length(levels(x))
q <- ................
m <- ................
V <- sqrt(....../(n * m))
return(....)
}
```

6. Calculer le coefficient de Cramer associé aux variables *Sexe* et *Diplome*.

Un autre exemple

Avec les lignes de code ci-dessous, on va générer aléatoirement (avec la fonction *sample*) 75 réalisations de deux variables qualitatives :

- type de film préféré : action ("A"), comédie ("C") et science-fiction ("S").
- niveau d'étude : collège ("C"), lycée ("L") et université ("U").

```
film <- as.factor(sample(c("A","C","S"),75,replace=T))
niveau<- as.factor(sample(c("C","L","U"),75,replace=T))
```

1. Construire le tableau de contingence.
2. A partir de cet échantillon, peut-on dire que le type du film préféré d'une personne dépend de son niveau d'étude ?
3. Retrouver la distance χ^2 du test précédent à partir des tableaux des effectifs observés et théoriques.
4. Calculer avec la fonction *cramer* (ci-dessus) le coefficient de cramer.