Chapitre 3

Tests du χ^2

3.1 Test du χ^2 d'indépendance

Petit rappel de cours : tableau de contingence, effectifs théoriques et observés, convergence en loi de la statistique de test, etc. (cf manuscript)

La fonction chisq.test
Cette commande travaille principalement avec le tableau de contingence que l'on souhaite tester. Ce dernier est passé en argument dans la fonction par l'intermédiaire d'une matrice.

```
> A <- matrix(c(12,22,22,25,55,12),2,3,byrow=TRUE)
> A
     [,1] [,2] [,3]
[1,]   12   22   22
[2,]   25   55   12
> chisq.test(A)

        Pearson's Chi-squared test

data:  A
X-squared = 13.7058, df = 2, p-value = 0.001056
```

En retour, on obtient :
- la valeur de la statistique de test (X-squared),
- le nombre de degrés de liberté (df pour degree of freedom),
- la p-valeur.

Au vu de ces informations, une décision peut immédiatement être prise. Il est cependant possible d'avoir accès à l'ensemble des calculs intermédiaires à l'aide des commandes :

```
> X <- chisq.test(A)
> X observed
     [,1] [,2] [,3]
[1,]   12   22   22
[2,]   25   55   12
> X expected
     [,1]     [,2]     [,3]
[1,]   14 29.13514 12.86486
[2,]   23 47.86486 21.13514
> X residuals
           [,1]      [,2]      [,3]
[1,] -0.5345225 -1.321885  2.546903
[2,]  0.4170288  1.031321 -1.987067
```

Dans cette situation, X représente une liste. On choisira observed pour les observations, expected pour les effectifs théoriques et residuals pour les résidus.

Quelques options additionnelles sont disponibles pour la fonction chisq.test :
- correct : Par défaut vaut FALSE. Dans le cas contraire, arrondie les valeurs théoriques à l'entier le plus proche dans le tableau de contingence.
- simulate.p.value : Si cette dernière vaut TRUE, indique que la p-valeur doit être approximée par la méthode de Monte-Carlo. Sinon, la p-valeur proposée et la plus proche pré-enregistrée : on perd donc en précision, ce qui peut nous amener à prendre une mauvaise décision.
- B : Dans le cas où simulate.p.value vaut TRUE, précise le nombre de réplications à utiliser pour approcher la p-valeur.

3.2 Test du χ^2 d'adéquation

Rappel de cours : distribution théorique connue ou non, lois continues et lois discrètes, etc (cf manuscript)

3.2.1 Lois discrètes

Comme pour le test du χ^2 d'indépendance, nous utiliserons ici la fonction chisq.test. La principale différence par rapport à la première partie réside dans les paramètres passés en argument. Cette fois ci on rentrera un vecteur x correspondant au fréquences observées pour les différentes classes et on précisera la probabilité théorique dans le champ p. Supposons que l'on dispose d'un échantillon et que l'on souhaite déterminer si ce dernier suit une loi théorique géométrique de paramètre $0,3$. La syntaxe à utiliser est alors :

```
> x <- c(40,29,8,10,13)
> proba <- c(0.3,0.21,0.147,0.103,0.240)
> chisq.test(x,p=proba)

        Chi-squared test for given probabilities

data:  x
X-squared = 14.4851, df = 4, p-value = 0.005897
```

Avec de tels résultats, on rejette l'hypothèse H_0 au niveau 0.01. La variable Y observée a peu de chance de suivre un loi géométrique. Comme pour le test d'indépendance, il est possible d'avoir accès au effectifs théoriques et au résidus :

```
> chisq.test(x,p=proba)$expected
[1] 30.0 21.0 14.7 10.3 24.0
> chisq.test(x,p=proba)$residuals
[1]  1.82574186  1.74574312 -1.74749578 -0.09347654 -2.24536560
```

Il est également d'approximer la p-valeur par la méthode de Monte-Carlo.

Remarque : Si la probabilité théorique des différentes classes est basée sur l'estimation de p paramètres, le nombre de degré de libertés de la statistique de test est diminué d'autant... ce qui n'est pas pris en compte par la fonction chisq.test. Il sera donc nécessaire de re-calculer *manuellement* la p-valeur.

3.2.2 Lois continues

Travailler avec des lois continues est un peu plus délicat dans la mesure où il n'existe pas de fonction dédiée. Il faut donc, préalablement à l'utilisation de la fonction chisq.test, découper l'échantillon en différentes classes et calculer pour chacune d'entres elles les probabilités associées (correspondant à des intervalles sur $[0,1]$) . Ce travail préliminaire peut-être réalisé à l'aide de la fonction qnorm (pour obtenir les quantiles). La fonction hist sera également très utile puisqu'elle calcule un certain nombre de quantités dont nous avons besoin.

Utilisation avancée de la fonction hist

La commande hist appliquée à un vecteur x crée en fait une liste contenant :

- breaks : les classes crées
- counts :nombre d'observations classe par classe
- density : valeur de la densité classe par classe
- etc.

Par exemple, pour avoir accès au découpage proposé par **R**, on tapera :

```
> hist(x)$breaks
```