TP5 : tests du Khi2

Lancez R et placez-vous dans **votre** dossier AD (avec Fichier →Changer le répertoire courant).

Rappel : tout ce que vous sauvegardez sur le disque C ou sur le "bureau" des machines en salle de TP est perdu dès que vous vous déconnectez.

Pour vérifier le nom du répertoire courant : `getwd()`.

Ouvrez un nouveau script et tapez toutes vos commandes dedans.

1 Test du χ^2 d'ajustement

On va reprendre l'exercice sur le dé vu en cours. Le fichier `N:l2ie/ad/de.txt` contient le résultat de 100 lancers d'un dé à 6 faces. Chargez ce fichier dans R.

On se demande si le dé est truqué ou pas. La fonction R qui permet d'effectuer les tests du χ^2 est `chisq.test` :

La syntaxe est : `chisq.test`($Vecteur_{EffectifsObserves}$ `, p=` $Vecteur_{ProbaTheoriques}$)

Le $Vecteur_{ProbaTheoriques}$ est le vecteur des probabilités attendues si le dé n'est pas truqué, c'est-à-dire ici $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$, et le $Vecteur_{EffectifsObserves}$ s'obtient avec la fonction `table` appliquée aux lancers (qui correspondent à nos observations).

Observez le résultat obtenu. On peut aussi faire afficher la valeur critique du χ^2 avec la commande `qchisq(0.95,df=5)` , où `df` est le nombre de degrés de liberté (ici nombre de modalités -1).

Comme pour les tests d'hypothèse qu'on verra par la suite, la `p-value` est le point important : on fait l'hypothèse (appelée H_0) que notre variable suit la loi théorique annoncée (ici la loi uniforme sur l'ensemble $\{1, 2, 3, 4, 5, 6\}$). La `p-value` donne la probabilité d'avoir cette valeur de d_{χ^2} (ici 16.04) si cette hypothèse est vraie. On peut aussi considérer qu'elle donne le risque qu'on prend à rejeter l'hypothèse H_0 si elle est vraie. Le risque maximum est généralement fixé à 5%, donc on va rejeter H_0 (c'est-à-dire déclarer que les lancers du dé ne suivent probablement pas une loi uniforme) si la `p-value` est $< 5\%$, et la conserver dans le cas contraire.

On a : (`p-value` $< 5\%$) $\iff$ ($d_{\chi^2} >$ seuil critique)
$\iff$ les observations ne suivent probablement pas la loi testée

Ici que peut-on dire ? (le dé est-il truqué ?)

NB : le test du $\chi 2$ ne renseigne pas sur les modalités responsables des écarts. En complément, on peut faire afficher la « contribution au χ^2 » de chaque terme avec la variable `residuals` : tapez `chisq.test(....,p=...)$residuals`. Le résultat obtenu est pour chaque « case » du tableau la valeur de $\frac{n_{obs} - n_{theo}}{\sqrt{n_{theo}}}$, ce qui permet de visualiser les modalités qui contribuent le plus au $\chi 2$ (pour préciser le résultat du test). Que peut-on dire ici ?

- - -

On peut faire ce type de test avec d'autres lois, en particulier binomiale, ou de Poisson. Pour les lois continues, il faut regrouper les données en classes.

Exercice avec ajustement à une loi de Poisson

Rappel : On dit qu'une variable aléatoire X suit une loi de Poisson de paramètre λ si

$$\forall k \in \mathbb{N}, P[X = k] = e^{-\lambda}.\frac{\lambda^k}{k!}$$

Elle modélise des événements qui se produisent avec une certaine régularité dans le temps.

Ici on a enregistré pendant une demi-journée le nombre X de clients entrant dans un magasin en une minute. On a obtenu le tableau suivant :

Nombre de clients k	0	1	2	3	4	5	plus de 5
Nombre de minutes où $X = k$	29	68	75	51	30	10	7

Avec un niveau de confiance de 95%, peut-on admettre que les arrivées suivent une loi de Poisson de paramètre $\lambda = 2$ clients par minute ?

2 Test du χ^2 d'indépendance

2.1 Données sous forme d'un tableau de contingence

On va reprendre l'exercice croisant dose de médicament reçue et guérison du cours. La saisie des données se fait de la façon suivante :

```
tab <- matrix(c(30,42,58,30,35,31), ncol=2)      # examinez ce qu'on a dans tab
rownames(tab)<-c("Dose D1","Dose D2","Dose D3")  # modalités des lignes
colnames(tab)<-c("Guéris","Non guéris")          # modalités des colonnes
tab                                              # visualisation des données
```

On peut éventuellement faire afficher les profils-lignes avec la fonction `prop.table` :

```
prop.table(tab,margin=1)                # mettre margin=2 pour les profils-colonnes
round(100*prop.table(tab,margin=1),1)  # autre affichage (pourcentage)
```

Observez les résultats produits.

Test du χ^2 :

On utilise comme précédemment la fonction `chisq.test` :

```
chisq.test(tab)                # observez
chisq.test(tab)$expected       # effectifs théoriques si indépendance totale
```

Faites afficher le seuil critique comme précédemment.
Quelle est la conclusion ?
Vous pouvez éventuellement faire afficher également les contributions au $\chi 2$ de chaque terme.

2.2 Données sous forme brute

Si les données sont sous forme brute (ce qui est généralement le cas après une enquête), il faut utiliser la fonction `table` pour obtenir les effectifs par modalités (le tableau de contingence).
Vous trouverez les données brutes correspondant à l'exemple précédent dans le fichier `medic.txt`.

1. Importez ce fichier de données, en appelant par exemple `medic` ces données. Examinez ce fichier.
2. Pour retrouver les tableaux de contingence, on utilise la fonction `table` :

   ```
   table(medic$Dose)       # effectifs par ligne
   table(medic$Effet)      # effectifs par colonne
   table(medic)            # tableau de contingence
   ```

   Faites afficher les effectifs théoriques en cas d'indépendance, faites le test du $\chi 2$ et concluez.

Exercice sur les données du Titanic :

Le Titanic est parti le 10 avril 1912 pour son premier voyage de Southampton en Angleterre vers New York. Le 14 avril il a heurté un iceberg et a coulé au fond de l'océan. Le nombre de personnes à bord est assez incertain (environ 2200), mais on sait que seules 771 personnes ont survécu. On dispose d'un fichier de données sur les passagers du Titanic (avec 2201 noms), comprenant pour chaque personne :

— une variable `Class` avec comme valeurs :
  - `first` pour un passager en première classe
  - `second` pour un passager en deuxième classe
  - `third` pour un passager en troisième classe
  - `crew` pour un membre d'équipage

— une variable `Age` avec comme valeurs `adult` ou `child`
— une variable `Sex` avec comme valeurs `male` ou `female`
— une variable `Survived` avec comme valeurs `yes` ou `no`

1. Importez le fichier de données `N:l2ie/ad/titanic.txt`, en appelant par exemple `tita` ces données. Examinez ce fichier.
2. Représentez les 4 variables par un diagramme en camembert (les 4 graphiques dans la même fenêtre).
3. On s'intéresse à la survie des passagers : on va croiser la variable `Survived` avec chacune des autres variables pour savoir dans quelle(s) catégorie(s) il valait mieux être pour avoir le plus de chances de survivre au naufrage.

   Répondez à ce problème en faisant des tests du $\chi 2$.

   Pour retrouver les tableaux de contingence, utilisez la fonction `table`. Quelques exemples :

```
table(tita$Survived)                          # effectifs liés à la variable survie
table(tita$Survived,tita$Age)                 # croisement de la survie et de l'âge
chisq.test(table(tita$Survived,tita$Age))     # test du khi2 entre survie et âge
chisq.test(table(tita$Survived,tita$Age))$residuals   # contributions
```

   Croisez aussi avec les deux autres variables et concluez.