Université de Montpellier
2018/2019

HLMA408: Traitement de données
Emmanuel Bonnet, Joseph Salmon

TP n° 2 : Tests d'adéquation, d'indépendance

Merci de lire la fiche de TP en détail. Avant d'appeler le responsable du TP, vérifier que vous avez bien lu la fiche en détail, et chercher dans les pages d'aide de R.

Les points à faire, ainsi que les questions attendant une réponse, sont marqués d'un triangle rouge ▸.

Les fichiers de données sont disponibles ici :

hcmv.data : http://josephsalmon.eu/enseignement/datasets/hcmv.data

babies23.data : http://josephsalmon.eu/enseignement/datasets/babies23.data

1 Test d'adéquation du χ^2 à une loi donnée

La fonction `chisq.test` permet de mettre en œuvre un test du χ^2 avec R. Pour plus de détail sur son fonctionnement, lire la page d'aide.

▸ Utilisez la commande `?chisq.test` et lisez l'aide

1.1 Loi discrète

Un couple de cobayes à pelage gris et lisse a donne naissance à 64 descendants. Leurs pelages se repartissent ainsi : 33 gris et lisses, 13 blancs et lisses, 15 gris et rudes et 3 blanc et rudes. Le modèle de Mendel donne les probabilités suivantes à chacun de ces cas (dans le même ordre) : $9/16, 3/16, 3/16, 1/16$.

▸ Saisir les données dans un vecteur ainsi `cobaye <-c(33, 13, 15, 3)`
▸ Saisir les probabilités théoriques ainsi `mendel <-c(9/16, 3/16, 3/16, 1/16)`
▸ Utiliser la commande `chisq.test` pour conclure.
▸ Pourquoi est-il marqué `"Warning message:In chisq.test..."` ?

1.2 Loi continue regroupée en classe

Au préalable, nous devons d'abord comprendre deux fonctions de R : `cut` et `table`. La fonction `cut` permet de faire du regroupement par classes d'intervalles. Voyons un exemple sur un échantillon simulé de 50 observations provenant de 50 variables aléatoires normales centrées réduites. Les classes sont $]-\infty, -2]$, $]-2, -1]$, $\ldots$, $]1, 2]$ et $]2, +\infty]$:

```
echantillon.sim <- rnorm(50)
bornes <- c(-Inf, -2, -1, 0, 1, 2, Inf)
echantillon.regroupe <- cut(echantillon.sim, breaks=bornes)
```

▸ Que fait la fonction `rnorm` ?

La fonction `table` calcule les effectifs (le nombre d'occurrences) de chaque valeur possible dans un échantillon. Si on reprend l'échantillon regroupé calculé ci-dessus, et que l'on applique la fonction `table`, on obtient les effectifs de chacune des classes :

```
effectifs.classes <- table(echantillon.regroupe)
```

On souhaite maintenant mettre en place un test du χ^2 pour vérifier que ces données simulées proviennent bien d'une loi normale centrée réduite.

La fonction `pnorm` permet de calculer les probabilités de la forme $P(Z \leqslant x)$ lorsque Z suit une loi normale centrée réduite et x est un nombre réel. Par exemple `pnorm(0)` donne 0.5 et `pnorm(1.96)` donne environ 0.975.

▸ Que vaut la probabilité de chacune des classes ci-dessus sous la loi normale centrée réduite ? Utilisez la fonction `pnorm`

▸ Que fait la fonction `diff` ? Et la commande ci-dessous ?

```
diff(pnorm(bornes)
```

▸ Faire un test du χ^2 pour regarder si l'on peut considérer que les effectifs observés des différentes classes proviennent d'un regroupement par classe d'une loi normale centrée réduite.

▸ Refaire tout ce qui précède avec un nouvel échantillon simulé de taille 1000 et les classes $]-\infty, -1.5]$, $]-1.5, -0.5]$, $]-0.5, 0.5]$, $]0.5, 1.5]$ et $]1.5, +\infty[$.

1.3 Un autre test de normalité (test de Shapiro-Wilk)

Le test de Shapiro-Wilk permet de décider entre

$$\mathcal{H}_0 \ : \ X \text{ suit une loi gaussienne} \quad \text{vs.} \quad \mathcal{H}_1 \ : \ X \text{ ne suit pas une loi gaussienne}$$

▸ Lire la page d'aide de la commande `shapiro.test` pour comprendre comment celle-ci fonctionne.

▸ Exécuter ce test sur les 100 données simulées du paragraphe 1.2 (`echantillon.sim`) et conclure.

▸ Exécuter ce test sur un échantillon de taille 100 tiré selon une loi exponentielle (d'espérance 1) et conclure.

2 Test d'adéquation du χ^2 à une famille de lois

Le logiciel R ne dispose pas de fonction toute faite pour répondre à ce problème. Pour cela, nous allons reprendre l'exemple de cours d'adéquation à la famille des lois de Poisson et mettre en place un test comparant

$\mathcal{H}_0$: X suit une loi de Poisson de paramètre λ inconnu
vs.
$\mathcal{H}_1$: X ne suit pas une loi de Poisson.

On rappelle que la variable X représente le nombre de palindromes dans une unité de longueur de 4000 pb (paires de bases).

▸ Importer et mettre en forme les données du fichier `hcmv.data` dans une table nommée `hcmv` dans R. Attention votre variable ne doit pas doit s'appeler `V1` mais `location`.

▸ Ce test suppose que l'on commence par estimer le paramètre λ de la loi de Poisson. Rappeler comment on fait. Calculer la valeur numérique sur le jeu de données et l'enregistrer dans une variable nommée `lambda.hat`

▸ Que font les commandes ci-dessous ?

```
dpois(3,lambda.hat)*58
# ou bien de façon équivalente
58*exp(-lambda.hat)*(lambda.hat)^3/factorial(3)
```

▸ Les commandes suivantes permettent de compter le nombre de palindromes dans chacune des régions de longueur 4000 paires de bases.

```
segments <- seq(from=1, to=232001, by=4000)
comptage.palin <- table(cut(hcmv$location, breaks=segments, labels=FALSE))
```

▸ Regrouper cette variable `comptage.palin` par classes dont les bornes sont $-\infty$, 2, 3, 4, 5, 6, 7, 8, $+\infty$.

▸ Que fait la commande ci-dessous ? Comparer avec la loi normale centrée réduite du paragraphe 1.2.

```
diff(ppois(classes, lambda.hat))
```

▸ On peut essayer de faire tourner la fonction `chisq.test` ici avec

```
test_chi2 <- chisq.test(effectifs.observes, p = prob.theo)
```

où `prob.theo` est le vecteur des probabilités "théoriques" de chaque classe.

▸ Que vous indique le message d'avis ? Comment feriez-vous pour le supprimer ?

▸ Refaire le test du χ^2 une fois que vous avez réglé le `"warning"` ci-dessus.

▸ Que signifie `df = 7` dans le texte fourni par R ? Est-ce en accord avec ce qu'il faudrait faire ? Pourquoi un tel problème ?

▸ Comment feriez-vous pour calculer la p-value correcte ? Les tables des lois du χ^2 s'obtiennent avec les commandes `pchisq` et `qchisq` et la commande `test_chi2$stat` donne la statistique de test. Conclure le test correctement.

3 Test d'indépendance du χ^2

Il existe un test d'indépendance du χ^2 que nous n'avons pas vu en cours ni en TD. Sont but est de discriminer entre les deux hypothèses suivantes :

$$\mathcal{H}_0 \ : \ X \text{ et } Y \text{ sont indépendantes} \quad \text{vs.} \quad \mathcal{H}_1 \ : \ X \text{ et } Y \text{ sont liées}$$

lorsque X et Y sont deux variables discrètes. Rappelons que deux variables sont dites indépendantes si connaître la valeur de l'une ne donne aucune information sur la valeur de l'autre. Par exemple, si on lance deux dés simultanément, l'un rouge, l'autre blanc et que l'on modélise par X le résultat du dé blanc et Y le résultat du dé rouge, alors X et Y sont indépendants.

Pour cela, on utilise la fonction `chisq.test` sur un tableau de contingence croisé entre deux variables. Il s'agit d'un tableau dont chaque ligne correspond à une valeur possible de X et chaque colonne à une valeur possible de Y. Dans la case sur la ligne x et la colonne y, on compte le nombre d'observations où l'on voit simultanément $X = x$ et $Y = y$. De tels tableaux de contingence se calculent avec la fonction `table`. Par exemple, sur les données de la table `babies` utilisée dans les précédents TP,

```
table(babies$ed, babies$smoke)
```

permet d'avoir un tableau de contingence croisant le niveau d'éducation des mères avec leur status tabagique.

▸ Reprendre la table `babies` de la fiche de TP1 (si vous avez enregistré le "workspace" à la fin de la fiche de TP1, il vous suffit de le recharger).

▸ Calculer le tableau de contingence croisé entre les variables `ed` et `smoke`. Que pensez-vous du résultat ? Enregistrer ce tableau dans un objet de R que l'on va appeler `tab.ed.smoke`

▸ Lancer `chisq.test(tab.ed.smoke)` et conclure. Pourquoi affiche-t-il un warning ?