Statistique Numérique et Analyse de Données
Ecole des Ponts ParisTech, 2ème année

TP : Test du χ^2

Rappels sur le test du χ^2 d'adéquation à une loi

On suppose qu'on dispose de n v.a.r. X_i discrètes (à valeurs dans $\{a_1, \ldots, a_k\}$) indépendantes et de même loi caractérisée par $\underline{p} = (p_1, \ldots, p_k)$, où $p_i = \mathbb{P}(X = a_i)$. On souhaite tester l'hypothèse que la loi est $\underline{p}^0 = (p_1^0, \ldots, p_k^0)$. On utilise la statistique suivante :

$$\xi_n = n \sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_i^0\right)^2}{p_i^0}.$$

Sous l'hypothèse nulle ($\underline{p} = \underline{p}^0$), ξ_n converge vers la loi du χ^2 à $k-1$ degrés de liberté, tandis qu'elle diverge sous l'hypothèse alternative ($\underline{p} \neq \underline{p}^0$). Au niveau de test α, on rejette l'hypothèse nulle si :

$$\xi_n \geq \chi^2_{k-1,1-\alpha}$$

où $\chi^2_{k-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $k-1$ degrés de liberté.

Exercice 1. Naissances et type de jour

Le tableau suivant contient les effectifs moyens hebdomadaires des naissances dans un hôpital américain pour l'année 1997 :

Type de jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Effectifs moyens	564772	629408	609596	604812	605280	450840	404456

En utilisant le test du χ^2 d'adéquation à une loi, on souhaite déterminer si les naissances sont réparties équitablement sur les jours de la semaine.

1. Saisissez les effectifs :

   ```
   effectif <- 52*c(564772,629408,609596,604812,605280,450840,404456)
   ```

2. Saisissez le vecteur $\underline{p}^0$:

   ```
   proba <- rep(1,7)/7
   ```

3. Effectuez le test du χ^2 :

   ```
   chisq.test(effectif,p=proba)
   ```

   ▶ Quelle conclusion tirez-vous ?

Exercice 2. Dés de Weldon

Weldon a réalisé $n = 26306$ lancers de 12 dés à 6 faces. Soit X_i la v.a.r. à valeurs dans $\{0, \ldots, 12\}$ désignant le nombre de faces comportant un cinq ou un six lors du i-ème lancer. Le tableau suivant contient les effectifs empiriques observées :

Nombre de faces 5 ou 6	0	1	2	3	4	5	6	7	8	9	10	11	12
Effectifs	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0

En utilisant le test du χ^2 d'adéquation à une loi, on souhaite déterminer si les dés sont équilibrés.

1. Saisissez les effectifs :

   effectif <-c (185,1149,3265,5475,6114,5194,3067,1331,403,105,14,4,0)

2. Quelle loi suit la v.a.r. X_i si les dés sont équilibrés ?
3. Saisissez le vecteur $\underline{p}^0$:

   proba <- dbinom(0 :12,12,1/3)

4. Effectuez le test du χ^2 :

   chisq.test(effectif,p=proba)

   ▶ Quelle conclusion tirez-vous ?

Exercice 3. Loi de Gompertz

Pour vérifier la loi de Gompertz modélisant l'extinction des populations, K. Miescher a observé en 1953 la durée de vie (exprimée en mois) de 144 rats d'un élevage :

Durée de vie	[10,15[	[15,20[	[20,25[	[25,28[	[28,30[	[30,32[
Effectifs	1	3	9	12	13	20

Durée de vie	[32,34[	[34,36[	[36,38[	[38,40[	[40,42[	[42,43[
Effectifs	23	26	22	11	3	1

La moyenne de l'échantillon est de 34.6 et l'écart-type de 5.3.
En utilisant le test du χ^2 d'adéquation à une loi, ont souhaite vérifier que le décalage de l'histogramme relativement à une courbe de loi normale $\mathcal{N}(32, 5)$ est significatif.

1. Saisissez les effectifs :

   effectif <- c(1,3,9,12,13,20,23,26,22,11,3,1)

   classe <- c(10,15,20,25,28,30,32,34,36,38,40,42,43)

2. Saisissez le vecteur $\underline{p}^0$:

   proba <- c()

   for (i in 1 :length(classe)-1) proba[i] <- pnorm(classe[i+1],mean=32,sd=5)-pnorm(classe[i],mean=32,sd=5)

   ▶ Commentez l'aspect "discret" de ce cas ?

3. Effectuez le test du χ^2 :

   chisq.test(effectif,p=proba,rescale.p=TRUE)

- ▶ Quelle conclusion tirez-vous ? Pourquoi utilise-on l'option `rescale.p=TRUE` ?

4. Effectuez de nouveau ce test du χ^2 en agrégeant certaines classes aux extrémités.
   - ▶ Commentez les résultats obtenus.

Rappels sur le test du χ^2 d'adéquation à une famille de lois discrètes

On souhaite maintenant tester l'hypothèse que la loi appartient à une famille $(\underline{p}_\theta)_{\theta\in\Theta}$ avec $\underline{p}_\theta = (p_{1,\theta}, \ldots, p_{k,\theta})$. L'idée naturelle consiste à remplacer dans le test précédent $\underline{p}^0$ par $\underline{p}_{\hat{\theta}}$, où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ. On démontre alors que si l'ensemble Θ des valeurs possibles pour θ est une partie ouverte d'intérieur non vide de $\mathbb{R}^h$ (avec $h < k-1$) alors la statistique :

$$\xi_n = n \sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_{i,\hat{\theta}}\right)^2}{p_{i,\hat{\theta}}}$$

converge vers la loi du χ^2 à $k-h-1$ degrés de liberté sous l'hypothèse nulle.

Exercice 4. Nombre de décès dans la cavalerie de l'armée prussienne

Le tableau suivant recense le nombre de décès dans la cavalerie prussienne sur la période 1875-1894, par année et par régiment, causés par un coup de sabot de cheval :

Nombre de décès par an et par régiment	0	1	2	3	4
Effectifs	144	91	32	11	2

En utilisant le test du χ^2 d'adéquation à une famille de lois, on souhaite vérifier si les données peuvent être modélisées par une loi de Poisson.

1. Saisissez les données :
   ```
   effectif <- c(144,91,32,11,2)
   ```
2. Déterminez le paramètre (potentiel) de la loi de Poisson par maximum de vraisemblance.
   ```
   x <- c(0,1,2,3,4)
   lambda <- weighted.mean(x,effectif/sum(effectif))
   ```
   - ▶ Vérifiez que l'estimateur du maximum de vraisemblance est la moyenne empirique.
3. Saisissez le vecteur $\underline{p}^0$:
   ```
   proba <- dpois(0 :4,lambda)
   ```
4. Effectuez le test du χ^2 :
   ```
   chisq.test(effectif,p=proba,rescale.p=TRUE)
   ```
   - ▶ Quelle conclusion tirez-vous ?

Rappels sur le test d'indépendance du χ^2

On considère maintenant un échantillon (X_i, Y_i) à valeurs dans $\{a_1, \ldots, a_k\} \times \{b_1, \ldots, b_l\}$. On veut tester indépendance des deux composantes X et Y (hypothèse nulle). En notant $p_{i,j} = \mathbb{P}\left((X,Y) = (a_i, b_j)\right)$, $q_i = \sum_{j=1}^{l} p_{i,j}$ et $r_j = \sum_{i=1}^{k} p_{i,j}$, on sait que la statistique :

$$\xi_n = n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(\hat{p}_{i,j} - \hat{q}_i \hat{r}_j\right)^2}{\hat{q}_i \hat{r}_j}$$

converge vers la loi du χ^2 à $(k-1)(l-1)$ degrés de liberté sous l'hypothèse nulle.

Exercice 5. Données de Fisher sur la couleur de cheveux

Le tableau suivant recense la couleur des cheveux de garçons et de filles d'un district écossais :

Couleur des cheveux	Blond	Roux	Châtain	Brun	Noir de jais
Nombre de garçons	592	119	849	504	36
Nombre de filles	544	97	677	451	14

En utilisant le test d'indépendance du χ^2, on souhaite déterminer si la couleur des cheveux est indépendante du sexe.

1. Saisissez les données :

```
effectif <- matrix(c(592,544,119,97,849,677,504,451,36,14),ncol=5)
rownames(effectif) <- c("Masculin","Féminin")
colnames(effectif) <- c("Blond","Roux","Châtain","Brun","Noir de jais")
```

2. Visualisez les données par des diagrammes en tuyaux d'orgue en fonction du sexe.

```
par(mfrow=c(2,1))
barplot(effectif[1,],main="Masculin")
barplot(effectif[2,],main="Féminin")
```

   ▶ Que pouvez-vous en dire ?

3. Effectuez le test du χ^2 :

```
khi2 <- chisq.test(effectif)
val_critique <- qchisq(0.95,df=3)
p_valeur <- 1-pchisq(khi2$statistic,df=3)
```

   ▶ Quelle conclusion pouvez-vous en tirer ? Justifier les calculs de `val_critique` et `p_valeur`

4. Calculez les différentes contributions au test du χ^2.

```
round(100*khi2$residuals^2/khi2$stat,1)
```

   ▶ Commentez.

5. Auriez-vous pu imaginer ce résultat à partir des profils ligne et colonne ?

```
round(100*prop.table(effectif,margin=1),1)
round(100*prop.table(effectif,margin=2),1)
```