Statistique - TD4 - Test du χ^2

Exercice 1.

1. Représentez la densité d'une variable aléatoire distribuée selon χ^2_{20}.

   R : `curve(dchisq(x,20),xlim=c(0,50))`

2. Avec quelle probabilité cette variable est-elle inférieure à 10 ?

   $\mathbb{P}(\chi^2_{20} \leq 10) \approx 0,032$. (R : `pchisq(10,20)`). Une variable aléatoire distribuée selon χ^2_{20} est inférieure ou égale à 10 dans $3,2\%$ des cas.

3. Déterminez et interprétez le quantile d'ordre $0,95$ de la loi χ^2_{20}.

   R : `qchisq(0.95,20)` donne environ : $31,4$. Le quantile d'ordre $0,95$ de la loi χ^2_{20} vaut $31,4$; une variable aléatoire distribuée selon χ^2_{20} est inférieure ou égale à $31,4$ avec 95% de chances.

Exercice 2.

Table 1 donne la répartition de soixante consommateurs de produits Danone par niveau de satisfaction (A/B/C/D) et par genre (H/F).

		satisfaction			
		A	B	C	D
genre	H	5	15	5	5
	F	5	10	5	10

Table 1: *Répartition observée de soixante sujets par rhésus et par groupe sanguin*

On souhaite tester au seuil de 5% l'indépendance du genre et du niveau de satisfaction d'un sujet choisi au hasard.

1. Enoncez les hypothèses $\mathscr{H}_0$ et $\mathscr{H}_1$.

   $\mathscr{H}_0$: la satisfaction est indépendante du genre ; $\mathscr{H}_1$: la satisfaction et le genre sont des variables dépendantes.

2. Sur quelle statistique la décision du test repose-t-elle ?

   La statistique de décision est : $X = \sum_{i=1}^{2}\sum_{j=1}^{4}(n_{i,j} - n^*_{i,j})^2/n^*_{i,j}$ avec :

   - $n_{i,j}$: effectif observé dans la modalité i de la variable *genre* et j de la variable *satisfaction*

   - $n^*_{i,j}$: effectif théorique dans la modalité i de la variable *genre* et j de la variable *satisfaction*.

   Pour rappel :

   n : taille de l'échantillon.

   $n_{i,\bullet} = \sum_{j=1}^{4} n_{i,j}$ effectif marginal dans la modalité i du genre.

   $n_{\bullet,j} = \sum_{i=1}^{2} n_{i,j}$ effectif marginal dans la modalité j de la satisfaction.

   $n^*_{i,j} = n_{i,\bullet} \times n_{\bullet,j}/n$.

3. Quelle est la distribution de la statistique de test sous $\mathscr{H}_0$?

   Sous $\mathscr{H}_0$, X est distribuée selon la loi de χ^2 à $(2-1)\times(4-1) = 3$ degrés de liberté.

4. Dressez le tableau des effectifs théoriques sous hypothèse d'indépendance.

   Les effectifs théoriques (Table 2) peuvent être calculés à la main : $n^*_{1,1} = n_{1,\bullet} \times n_{\bullet,1}/n = 30 \times 10/60$, etc. On peut aussi les obtenir sous R :

   - effectifs observés : `cont = matrix(c(5,15,5,5,5,10,5,10),nrow=2,byrow=TRUE)`

   - effectifs théoriques : `chisq.test(cont)$expected`

5. Quelle est la valeur observée de la statistique de test ?

   La valeur observée de statistique de test est : $X_{obs} = (5-5)^2/5 + (15-12,5)^2/12,5 + \cdots + (10-7,5)^2/7,5 = 8/3 \approx 2,67$. On l'obtient sous R par : `theo = chisq.test(cont)$expected ; (cont-theo)^2/theo` ou bien : `chisq.test(cont)$statistic`.

A. Lourme, Faculté d'économie, gestion & AES, Université de Bordeaux `http://alexandrelourme.free.fr`

		satisfaction				
		A	B	C	D	
genre	H	$n^*_{1,1} = 5$	$n^*_{1,2} = 12,5$	$n^*_{1,3} = 5$	$n^*_{1,4} = 7,5$	$n_{1,\bullet} = 30$
	F	$n^*_{2,1} = 5$	$n^*_{2,2} = 12,5$	$n^*_{2,3} = 5$	$n^*_{2,4} = 7,5$	$n_{2,\bullet} = 30$
		$n_{\bullet,1} = 10$	$n_{\bullet,2} = 25$	$n_{\bullet,3} = 10$	$n_{\bullet,4} = 15$	$n = 60$

Table 2: *Tableau des effectifs théoriques sous hypothèse d'indépendance*

6. Quelle est la valeur critique du test ?

   Puisque le seuil de significativité est 5%, la valeur critique est le quantile d'ordre $0,95$ de χ^2_3 : $7,8$ environ.

7. Déterminez la p-valeur du test.

   La p-valeur est la probabilité que sous $\mathscr{H}_0$, la statistique X prenne une valeur plus atypique encore que la valeur observée : p-val. $= \mathbb{P}(X > X_{obs}) = \mathbb{P}(\chi^2_3 > 8/3) = 1 - \mathbb{P}(\chi^2_3 \leq 8/3) \overset{*}{\approx} 0,446$ ($*$ R : `1-pchisq(8/3,3)`).

8. Enoncez la décision du test de deux façons :

   i. en comparant la valeur observée de la statistique de test à la valeur critique.
   La valeur observée de X ($X_{obs} = 8/3$) est inférieure à la valeur critique ; au seuil 5% on ne rejette pas l'indépendance des variables *genre* et *satisfaction*.

   ii. en comparant la p-valeur au seuil de risque.
   La p-valeur est supérieure au seuil de significativité (5%) ; au seuil 5% on ne rejette pas l'indépendance des variables *genre* et *satisfaction*.

Exercice 3

Table 3 consigne les résultats d'une enquête : on demande à deux-cents personnes nées en Aquitaine si elles sont favorables à la radiation de la corrida du patrimoine culturel immatériel (PCI).

		Radiation		
		oui	neutre	non
Département	Dordogne	15	14	21
	Gironde	19	14	8
	Landes	10	13	10
	Lot-et-Garonne	14	12	17
	Pyrénées-Atlantiques	11	13	9

Table 3: *Répartition de deux-cents personnes selon : (i) leur département d'origine et (ii) leur avis sur la radiation de la corrida du PCI*

On souhaite déterminer dans un test de χ^2 si *Radiation* et *Département* sont des variables indépendantes.
$\mathscr{H}_0$: *Radiation* et *Département* sont des variables indépendantes ; $\mathscr{H}_1$: *Radiation* et *Département* ne sont pas indépendantes.

1. Dressez le tableau des effectifs théoriques sous hypothèse d'indépendance.

   Les effectifs théoriques sont dans Table 4.

		Radiation		
		oui	neutre	non
Département	Dordogne	$17,2$	$16,5$	$16,2$
	Gironde	$14,1$	$13,5$	$13,3$
	Landes	$11,4$	$10,9$	$10,7$
	Lot-et-Garonne	$14,8$	$14,2$	$14,0$
	Pyrénées-Atlantiques	$11,4$	$10,9$	$10,7$

Table 4: *Tableau des effectifs théoriques (arrondis) sous hypothèse d'indépendance*

2. Déterminez la valeur de la statistique du χ^2.

   La valeur observée de la statistique de test : $X_{obs} \approx 8,24$

3. Doit-on rejeter l'indépendance des deux variables au seuil de 5% ?

   La valeur critique associée au seuil de significativité 5% est le quantile d'ordre $0,95$ de la loi de χ^2 à $(5-1)\times(3-1)=8$ degrés de liberté : $15,51$. Puisque la valeur critique est supérieure à la valeur observée de la statistique de test, on ne rejette pas l'indépendance entre la région d'origine et l'avis d'un sujet sur la radiation de la corrida du PCI.

4. Déterminez et interprétez la p valeur du test.

   La p-valeur du test est la probabilité pour une variable distribuée selon χ^2_8 d'être supérieure à $X_{obs} \approx 8,24$; elle vaut environ $0,41$. Si le seuil de significativité est inférieur à 41%, on ne rejette pas l'indépendance.

Exercice 4.

Table 5 répartit cent téléspectateurs selon le programme préféré et la chaîne favorite.

		programme préféré		
		film	journal	sport
chaîne favorite	Ch. 1	12	13	9
	Ch. 2	10	18	11
	Ch. 3	7	12	8

Table 5: *Répartition de cent téléspectateurs selon le programme préféré et la chaîne favorite*

On souhaite déterminer dans un test de χ^2 si le programme et la chaîne préférés d'un téléspectateur sont indépendants. $\mathscr{H}_0$: *programme préféré* et *chaîne favorite* sont des variables indépendantes ; $\mathscr{H}_1$: *programme préféré* et *chaîne favorite* ne sont pas indépendantes.

1. Dressez le tableau des effectifs théoriques sous hypothèse d'indépendance.

		programme préféré		
		film	journal	sport
chaîne favorite	Ch. 1	$9,86$	$14,62$	$9,52$
	Ch. 2	$11,31$	$16,77$	$10,92$
	Ch. 3	$7,83$	$11,61$	$7,56$

Table 6: *Effectifs théoriques sous hypothèse d'indépendance*

2. Déterminez la valeur de la statistique du χ^2.

   $X_{obs} \approx 1,04$.

3. Doit-on rejeter l'indépendance des deux variables au seuil de 5% ?

   p-val. $= \mathbb{P}(\chi^2_4 > X_{obs}) = 1 - \mathbb{P}(\chi^2_4 \leq X_{obs}) \approx 0,90$. p-val. $\geq \alpha = 0,05$; on ne rejette pas $\mathscr{H}_0$.

Exercice 5.

On considère le jeu de données HairEyeColor de R.

1. La couleur des yeux est-elle indépendante de celle des cheveux au seuil de 5% ?

   `HairEyeColor[„1]` # tableau de contingence (`couleur des cheveux` × `couleur des yeux`) chez les hommes

   `HairEyeColor[„2]` # tableau de contingence chez les femmes

   `cont = HairEyeColor[„1] + HairEyeColor[„2]` # tableau de contingence marginalement au sexe.

   `mytest = chisq.test(cont)` # le test du χ^2

   `mytest$p.value` # p-valeur du test

   La p-valeur du test ($\approx 2,3\times 10^{-25}$) est inférieure au seuil de significativité $(0,05)$; on rejette l'indépendance de la couleur des yeux et des cheveux.

2. La couleur des yeux des garçons est-elle indépendante de celle de leurs cheveux ?

   `mytest = chisq.test(HairEyeColor[„1])` # le test du χ^2 sur le tableau de contingence des hommes

   `mytest$p.value` # p-valeur du test

   La p-valeur du test ($\approx 4,4\times 10^{-6}$) est inférieure au seuil de significativité $(0,05)$; on rejette l'indépendance de la couleur des yeux et des cheveux chez les hommes.

Exercice 6.

On considère le jeu de données Ronfle[a] de Gilles Hunault. Dans la population étudiée, le tabagisme et le sexe sont-ils

[a] http://forge.info.univ-angers.fr/ gh/Datasets/ronfle.htm

indépendants ?

Après avoir enregistré les données : `http://forge.info.univ-angers.fr/~gh/Datasets/ronfle.htm` dans un dossier `myfolder` sous le nom `ronfle` :

```
whole <- read.table(file='myfolder/ronfle',sep='',header=TRUE) # lecture du fichier ronfle
mydata <- table(whole[,c(6,8)]) # tableau de contingence : sexe × tabac
mytest <- chisq.test(mydata) # test du χ²
mytest$expected # effectifs théoriques sous hypothèse d'indépendance
mytest$statistic # valeur observée de la statistique de test
mytest$p.value # p-valeur associée à la valeur observée de la statistique de test.
```

La p-valeur associée à la valeur observée de la statistique de test est : $0,008$. Les variables `sexe` et `tabac` sont jugées indépendantes quand le seuil de significativité est inférieur à $0,8\%$.

Exercice 7.

On considère le jeu de données Automobile[b] de l'University of California Irvine. En deçà de quelle valeur du seuil de risque le mode d'aspiration (standard/turbo) est-il jugé indépendant du type de carburant (fuel/gas) ?

```
whole <- read.table(file='http://archive.ics.uci.edu/ml/machine-learning-databases/
autos/imports-85.data',sep=',',header=FALSE) # lecture du fichier complet
mydata <- table(whole[,c(4,5)]) # tableau de contingence : fuel-type × aspiration
mytest <- chisq.test(mydata) # test du χ²
mytest$expected # effectifs théoriques sous hypothèse d'indépendance
mytest$statistic # valeur observée de la statistique de test
mytest$p.value # p-valeur du test
```

La p-valeur associée à la valeur observée de la statistique de test est : $5,3 \times 10^{-8}$. Les variables `fuel-type` et `aspiration` sont jugées indépendantes quand le seuil de significativité est inférieur à $5,3 \times 10^{-8}$.

Rappels de cours

Un échantillon (aléatoire) de taille n est décrit par deux variables qualitatives. La première X possède r modalités : $x_1, \ldots, x_r$; la seconde Y possède s modalités : $y_1, \ldots, y_s$. Pour $i \in \{1, \ldots, r\}$ et $j \in \{1, \ldots, s\}$ on note $n_{i,j}$ le nombre d'unités pour lesquelles X vaut x_i et Y vaut y_j, $n_{i,\bullet} = \sum_{l=1}^{s} n_{i,l}$ le nombre d'unités pour lesquelles X vaut x_i et $n_{\bullet,j} = \sum_{k=1}^{r} n_{k,j}$ le nombre d'unités pour lesquelles Y vaut y_j. Table 7 est un tableau de contingence qui regroupe les effectifs conjoints observés $n_{i,j}$ ainsi que les effectifs marginaux $n_{\bullet,j}$ et $n_{i,\bullet}$.

		Y				
		y_1	y_2	$\ldots$	y_s	*marge X*
X	x_1	$n_{1,1}$	$n_{1,2}$	$\ldots$	$n_{1,s}$	$n_{1,\bullet}$
	x_2	$n_{2,1}$	$n_{2,2}$	$\ldots$	$n_{2,s}$	$n_{2,\bullet}$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	x_r	$n_{r,1}$	$n_{r,2}$	$\ldots$	$n_{r,s}$	$n_{r,\bullet}$
	marge Y	$n_{\bullet,1}$	$n_{\bullet,2}$	$\ldots$	$n_{\bullet,s}$	n

Table 7: *Tableau de contingence : effectifs conjoints observés et effectifs marginaux*

Si les variables X et Y sont indépendantes, en négligeant l'effet d'échantillonnage le nombre d'individus pour lesquels $X = x_i$ et $Y = y_j$ est $n^*_{i,j} = n_{i,\bullet} \times n_{\bullet,j}/n$. Les coefficients $n^*_{i,j}$ regroupés dans Table 8 sont les effectifs conjoints théoriques ou attendus sous hypothèse d'indépendance.

Toujours si X et Y sont indépendantes, la statistique $E = \sum_{i=1}^{r} \sum_{j=1}^{s} (n_{i,j} - n^*_{i,j})^2/n^*_{i,j}$ est approximativement distribuée selon $\chi^2_{(r-1)\times(s-1)}$, la loi de χ^2 à $(r-1) \times (l-1)$ degrés de liberté.

Ainsi, lorsque l'hypothèse $\mathscr{H}_0$: 'X et Y sont indépendantes' est testée au seuil α $(0 < \alpha < 1)$ contre $\mathscr{H}_1$: 'X et Y ne sont pas indépendantes', $\mathscr{H}_0$ est rejetée si E est supérieure au quantile d'ordre $1 - \alpha$ de la loi $\chi^2_{(r-1)\times(s-1)}$.

$\lambda_{j|i} = n_{i,j}/n_{i,\bullet}$ représente la proportion de sujets pour lesquels Y vaut y_j parmi ceux pour lesquels X vaut x_i et $\lambda_j = n_{\bullet,j}/n$ la proportion (non conditionnelle) d'unités pour lesquelles Y vaut y_j. Les distributions de Y observées conditionnellement aux modalités de X – que l'on appelle les profils lignes – et la distribution marginale de Y sont regroupées dans Table 9.

[b] http://archive.ics.uci.edu/ml/datasets/Automobile

		Y				
		y_1	y_2	$\dots$	y_s	*marge* X
X	x_1	$n^*_{1,1}$	$n^*_{1,2}$	$\dots$	$n^*_{1,s}$	$n_{1,\bullet}$
	x_2	$n^*_{2,1}$	$n^*_{2,2}$	$\dots$	$n^*_{2,s}$	$n_{2,\bullet}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	x_r	$n^*_{r,1}$	$n^*_{r,2}$	$\dots$	$n^*_{r,s}$	$n_{r,\bullet}$
	marge Y	$n_{\bullet,1}$	$n_{\bullet,2}$	$\dots$	$n_{\bullet,s}$	n

Table 8: *Tableau de contingence : effectifs conjoints attendus (ou théoriques) et effectifs marginaux*

		Y				
		y_1	y_2	$\dots$	y_s	
X	x_1	$\lambda_{1\|1}$	$\lambda_{2\|1}$	$\dots$	$\lambda_{s\|1}$	
	x_2	$\lambda_{1\|2}$	$\lambda_{2\|2}$	$\dots$	$\lambda_{s\|2}$	*distributions conditionelles*
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	*de* Y *(profils lignes)*
	x_r	$\lambda_{1\|r}$	$\lambda_{2\|r}$	$\dots$	$\lambda_{s\|r}$	
		λ_1	λ_2	$\dots$	λ_s	*distribution marginale de* Y

Table 9: *Distributions conditionnelles (profils lignes) et distribution marginale de* Y

Lorsque X et Y sont indépendantes, les r profils lignes sont identiques[a] et chacun d'eux est égal à la distribution marginale de Y: pour tout $j \in \{1, \dots, s\}$, $\lambda_{j|1} = \lambda_{j|2} = \dots = \lambda_{j|r} = \lambda_j$.

On peut aussi caractériser l'indépendance grâce aux distributions conditionnelles de X – ou profils colonnes – définies par : $\gamma_{i|j} = n_{i,j}/n_{\bullet,j}$ $(j \in \{1, \dots, s\}, i = 1, \dots, r)$ et à la distribution marginale de X définie par : $\gamma_i = n_{i,\bullet}/n$ $(i = 1, \dots, r)$. Lorsque X et Y sont indépendantes, les s profils colonnes sont identiques et chacun d'eux est égal à la distribution marginale de X: pour tout $i \in \{1, \dots, r\}$, $\gamma_{i|1} = \gamma_{i|2} = \dots = \gamma_{i|s} = \gamma_i$.

[a] en négligeant toujours l'effet d'échantillonnage