Tests du khi-deux

C. Maugis-Rabusseau
Bureau 116, GMM
cathy.maugis@insa-toulouse.fr

4MA, 2020-2021

Introduction

- La famille des tests du khi-deux regroupe des tests d'objectifs variés (ajustement, indépendance, homogénéité, ...)
- Ces tests ont en commun de mesurer l'écart à l'hypothèse nulle via une "divergence du khi-deux"
- Les statistiques de test associées suivent asymptotiquement une loi du khi-deux
- Les tests du khi-deux sont valables pour l'étude de données qualitatives (ou discrètes) à support fini.
  Cependant, en pratique, ces tests sont aussi appliqués à des données discrètes à support infini ou continues après regroupement en classes

Plan

Objectif et principe du test

- X v.a. qualitative ou quantitative discrète à $K > 1$ modalités $\{a_1, \ldots, a_K\}$, de loi $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ inconnue

$$\pi_k = \mathbb{P}(X = a_k) > 0, \ \forall k \in \{1, \ldots, K\}.$$

- n-échantillon $(X_1, \ldots X_n)$ de même loi que X
- Loi de probabilité $\mathcal{L}^0$ sur $\{a_1, \ldots, a_K\}$ connue :

$$\mathbf{p}^0 = (p_1^0, \ldots, p_K^0)$$

- On veut tester : $\mathcal{H}_0 : X \sim \mathcal{L}^0$ contre $\mathcal{H}_1 : X \nsim \mathcal{L}^0$

$$\mathcal{H}_0 : \forall k, \ \pi_k = p_k^0 \text{ contre } \mathcal{H}_1 : \exists k, \ \pi_k \neq p_k^0$$

Objectif et principe du test

- Idée : estimer la loi π de X à l'aide de l'échantillon $(X_1, \ldots, X_n)$ et comparer cet estimateur avec $\mathbf{p}^0$
- $\forall k$, on estime π_k par $\hat{\pi}_k = \frac{N_k}{n}$ avec

$$N_k = \sum_{i=1}^{n} \mathbb{1}_{X_i = a_k}$$

- Divergence du khi-deux entre les lois $\hat{\pi}$ et $\mathbf{p}^0$:

$$T_n = n \sum_{k=1}^{K} \frac{\left(\hat{\pi}_k - p_k^0\right)^2}{p_k^0} = \sum_{k=1}^{K} \frac{\left(N_k - np_k^0\right)^2}{np_k^0}$$

Lien avec la loi multinomiale

- $N = (N_1, \ldots, N_K)'$ suit une loi multinomiale $\mathcal{M}(n, \boldsymbol{\pi})$ sur $\mathbb{N}^K$: pour tout $(n_1, \ldots, n_K) \in \mathbb{N}^K$,

$$\mathbb{P}(N_1 = n_1, \ldots, N_K = n_K) = \begin{cases} \frac{n!}{n_1! \ldots n_K!} \pi_1^{n_1} \ldots \pi_K^{n_K} & \text{si } \sum_{k=1}^K n_k = n \\ 0 & \text{sinon.} \end{cases}$$

- Les hypothèses du test peuvent se traduire par

$$\mathcal{H}_0 : N \sim \mathcal{M}(n, \mathbf{p}^0) \text{ contre } \mathcal{H}_1 : N \text{ ne suit pas } \mathcal{M}(n, \mathbf{p}^0)$$

Lien avec la loi multinomiale

- Soit $\sqrt{\pi} = \left(\sqrt{\pi_1}, \ldots, \sqrt{\pi_K}\right)'$.

$$Y_n = \left(\frac{N_1 - n\pi_1}{\sqrt{n\pi_1}}, \ldots, \frac{N_K - n\pi_K}{\sqrt{n\pi_K}} \right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_K\left(0, \Gamma\right),$$

où $\Gamma = I_K - (\sqrt{\pi})(\sqrt{\pi})'$ = matrice de projection orthogonale sur $\text{Vect}(\sqrt{\pi})^{\perp}$.

- Sous l'hypothèse que $X_1, \ldots, X_n$ sont i.i.d. de loi $\pi = (\pi_1, \ldots, \pi_K)$,

$$Z_n = \sum_{k=1}^{K} \frac{(N_k - n\,\pi_k)^2}{n\,\pi_k} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K-1).$$

Procédure de test

- Statistique de test :

$$T_n = \sum_{k=1}^{K} \frac{\left(N_k - np_k^0\right)^2}{np_k^0}$$

- Région de rejet au niveau "asymptotique" α :

$$\mathcal{R}_\alpha = \left\{ T_n > x_{K-1,1-\alpha} \right\}$$

où $x_{K-1,1-\alpha}$ est le $1-\alpha$ quantile d'un $\chi^2(K-1)$

- Sous $\mathcal{H}_0$, $T_n \overset{\mathcal{L}}{\simeq} \chi^2(K-1)$ dès lors que $np_k^0 \geq 5$ pour tout k. Lorsque ce n'est pas le cas, on regroupe des classes jusqu'à ce que ces conditions soient vérifiées. Mais lorsqu'on regroupe des modalités, la région de rejet change car la loi limite dépend du nombres de modalités.

Procédure de test

Que peut-on dire de la puissance du test ?

On peut remarquer que

$$\frac{T_n}{n} \geq \|N/n - \mathbf{p}^0\|^2 \xrightarrow{p.s} \|\pi - \mathbf{p}^0\|^2$$

par la loi des grands nombres et donc $T_n \xrightarrow{p.s} +\infty$. La puissance du test tend donc vers 1 quand n tend vers $+\infty$.

Exemple de Mendel

Chez les pois, le caractère couleur est codé par un gène présentant deux formes allèles J et v correspondant aux couleurs jaune et vert. Le jaune est dominant et le vert récessif. Le caractère de forme, rond ou ridé, est porté par un autre gène à deux allèles R (dominant) et r (récessif). On croise 2 populations (pures) de pois : l'une jaune et ronde, l'autre verte et ridée. Selon la prédiction de Mendel, au bout de 2 croisements, la proportion de pois

- **JR** jaunes et ronds est $9/16$
- **Jr** jaunes et ridés est $3/16$
- **vR** verts et ronds est $3/16$
- **vr** verts et ridés est $1/16$

Dans ses expériences, Mendel a obtenu les résultats suivants : $N_{JR} = 315$, $N_{Jr} = 101$, $N_{vR} = 108$, $N_{vr} = 32$. Ici $K = 4$ et l'on obtient que $(T_n)^{obs} = 0.47$ et $x_{3,0.95} = 7.82$. On accepte donc très largement l'hypothèse de Mendel.

Plan

Principe du test

- Θ un ouvert de $\mathbb{R}^d$ avec $1 \leq d < K$
- $(\mathcal{L}(\theta))_{\theta \in \Theta}$ famille de lois de probabilités définies sur $\{a_1, \ldots, a_K\}$
- On veut tester

$$\mathcal{H}_0 : \exists \theta \in \Theta,\ X \sim \mathcal{L}(\theta)$$

contre

$$\mathcal{H}_1 : \text{la loi de } X \text{ n'appartient pas à } (\mathcal{L}(\theta))_{\theta \in \Theta}.$$

Principe du test

- Les lois $(\mathcal{L}(\theta))_{\theta\in\Theta}$ sont caractérisées par les vecteurs de probabilités sur $\{a_1, \ldots, a_K\}$

$$\mathcal{P}(\Theta) = \{\mathbf{p}(\theta) = (p_1(\theta), \ldots, p_K(\theta));\ \theta \in \Theta\}\,.$$

- On souhaite donc tester

$$\mathcal{H}_0 : \boldsymbol{\pi} \in \mathcal{P}(\Theta) \text{ contre } \mathcal{H}_1 : \boldsymbol{\pi} \notin \mathcal{P}(\Theta).$$

- Idée du test : remplacer $\mathbf{p}^0$ dans T_n par la loi de $\mathcal{P}(\Theta)$ "la plus proche" de π au vu des données, c'est-à-dire la loi $\mathbf{p}(\hat{\theta})$ où $\hat{\theta}$ est l'EMV de θ basé sur $(X_1, \ldots, X_n)$ sous $\mathcal{H}_0$.

Procédure de test

- Statistique de test :

$$\widehat{T}_n = n \sum_{k=1}^{K} \frac{\left(\hat{\pi}_k - p_k(\hat{\theta})\right)^2}{p_k(\hat{\theta})} = \sum_{k=1}^{K} \frac{\left(N_k - np_k(\hat{\theta})\right)^2}{np_k(\hat{\theta})}$$

- Par proposition admise du cours .... sous $\mathcal{H}_0$,

$$\widehat{T}_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K-d-1).$$

- Région de rejet au niveau α :

$$\mathcal{R}_\alpha = \left\{ \widehat{T}_n > x_{K-d-1,1-\alpha} \right\}$$

où $x_{K-d-1,1-\alpha}$ est le $1-\alpha$ quantile de la loi $\chi^2(K-d-1)$.

- La p-valeur vaut

$$p((\widehat{T}_n)^{obs}) = \mathbb{P}_{\mathcal{H}_0}(\widehat{T}_n \geq (\widehat{T}_n)^{obs}) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(\chi^2(K-d-1) \geq (\widehat{T}_n)^{obs}).$$

Exemple

Pour 10000 fratries de 4 enfants, on a relevé le nombre de garçons :

Nb de garçons (k)	0	1	2	3	4
Effectifs (N_k)	572	2329	3758	2632	709

On décide de modéliser les naissances en supposant qu'elles sont indépendantes et que la probabilité d'avoir un garçon vaut $\theta \in]0,1[$.
X_i = nombre de garçons dans la $i$ème fratrie.

$$\mathcal{H}_0 : X_i \sim \mathcal{B}in(4,\theta) \text{ contre } \mathcal{H}_1 : X_i \nsim \mathcal{B}in(4,\theta), \theta \in]0,1[.$$

Sous $\mathcal{H}_0$, $\hat{\theta}_{MV} = \frac{\overline{X}_n}{4}$ et

$$\hat{T}_n = \sum_{k=0}^{4} \frac{\left(N_k - np_k(\hat{\theta})\right)^2}{np_k(\hat{\theta})} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(5-1-1) = \chi^2(3)$$

avec $p_k(\hat{\theta}) = \mathbb{P}(U = k)$ pour $U \sim \mathcal{B}in(4,\hat{\theta})$.

Exemple

```
> classes = c(0,1,2,3,4)
> Nk = c(572,2329,3758,2632,709)
> n = sum(Nk)
> pihat = Nk / n
> thetahat = sum(Nk * classes) / (n*4)
> ptheo = dbinom(0:4,4,thetahat)
> Tobs = sum(((Nk - (n*ptheo))^2) / (n*ptheo))
> print(Tobs)
[1] 0.9882779
> val = chisq.test(Nk,p=ptheo)  # Attention aux degrés de liberté !
> print(val)

Chi-squared test for given probabilities

data:  Nk
X-squared = 0.98828, df = 4, p-value = 0.9116

> pval = 1 - pchisq(val$statistic,3)
> print(pval)
X-squared
0.8040883
```

Plan

Objectif

- X v.a avec K modalités $\{a_1, \ldots, a_K\}$
- Y v.a avec L modalités $\{b_1, \ldots, b_L\}$
- Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendants et de même loi que (X, Y)

- On souhaite tester

$$\mathcal{H}_0 : X \text{ et } Y \text{ sont indépendantes}$$

contre

$$\mathcal{H}_1 : X \text{ et } Y \text{ ne sont pas indépendantes.}$$

Idée du test

- La loi du couple (X, Y) est caractérisée par les probabilités

$$\mathbb{P}(X = a_k, Y = b_l) \text{ pour tout } k = 1, \ldots, K, \ l = 1, \ldots, L.$$

- Sous $\mathcal{H}_0$: $\forall (k, l), \ \mathbb{P}(X = a_k, Y = b_l) = \mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$
- Sous $\mathcal{H}_1$: $\exists (k, l), \ \mathbb{P}(X = a_k, Y = b_l) \neq \mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$.
- On estime
  - $\mathbb{P}(X = a_k, Y = b_l)$ par $\frac{N_{k,l}}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = a_k, Y_i = b_l\}}$
  - $\mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$ par $\frac{N_{k,.} N_{.,l}}{n^2} = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{1}_{X_i = a_k} \sum_{i=1}^{n} \mathbb{1}_{Y_i = b_l}$
- On obtient la statistique de test :

$$I_n = n \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(\frac{N_{k,l}}{n} - \frac{N_{k,.} N_{.,l}}{n^2} \right)^2}{\frac{N_{k,.} N_{.,l}}{n^2}} = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n} \right)^2}{\frac{N_{k,.} N_{.,l}}{n}}$$

Procédure de test

- Statistique de test :

$$I_n = n \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(\frac{N_{k,l}}{n} - \frac{N_{k,.} N_{.,l}}{n^2} \right)^2}{\frac{N_{k,.} N_{.,l}}{n^2}} = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n} \right)^2}{\frac{N_{k,.} N_{.,l}}{n}}$$

- On suppose que $\forall k,\ \mathbb{P}(X = a_k) > 0$ et $\forall l,\ \mathbb{P}(Y = b_l) > 0$. Alors, sous $\mathcal{H}_0$,

$$I_n \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2((K-1)(L-1))$$

- Région de rejet au niveau "asymptotique" α :

$$\mathcal{R}_\alpha = \{ I_n > x_{(K-1)(L-1),1-\alpha} \}$$

Exemple

Une enquête a été réalisée auprès d'un échantillon de 250 personnes au sujet de l'abaissement à 16 ans du droit de vote. Les réponses ont été classées suivant le niveau d'instruction des personnes interrogées :

Niveau d'instruction	Pour	Contre	$N_{k.}$
Brevet	10	15	25
Bac	20	85	105
Bac +2 et plus	20	100	120
$N_{.l}$	50	200	250

```
> contingence = matrix(c(10,20,20,15,85,100),ncol=2)
> chisq.test(contingence)

Pearson's Chi-squared test

data:  contingence
X-squared = 7, df = 2, p-value = 0.03
```

Plan

Test d'homogénéité

- On étudie un caractère pouvant prendre K valeurs : $\{a_1, \ldots, a_K\}$.
- On dispose de L échantillons $E_1, \ldots, E_L$
- N_{kl} = effectif observé de la valeur a_k dans l'échantillon E_l
- $N_{k.} = \sum_{l=1}^{L} N_{kl}$, $N_{.l} = \sum_{k=1}^{K} N_{kl}$, $n = \sum_{k=1}^{K} \sum_{l=1}^{L} N_{kl}$
- On veut tester

$$\mathcal{H}_0 : \text{ les éch. sont issus de la même loi}$$

contre

$$\mathcal{H}_0 : \text{ les éch. ne sont pas issus de la même loi}$$

Test d'homogénéité

- Statistique de test :

$$J_n = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.} N_{.,l}}{n} \right)^2}{\frac{N_{k,.} N_{.,l}}{n}}$$

- Sous $\mathcal{H}_0$, $J_n \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2((K-1)(L-1))$

- Région de rejet au niveau "asymptotique" α :

$$\mathcal{R}_\alpha = \{J_n > x_{(K-1)(L-1),1-\alpha}\}$$

Exemple

- 2 populations : collège A et collège B
- X= participation à un club sportif $\rightarrow \{a_1, a_2\} = \{"oui", "non"\}$
- On veut tester

$$\mathcal{H}_0 : \text{ les 2 populations sont homogènes (même taux de participation)}$$

contre

$$\mathcal{H}_1 : \text{ les 2 populations ne sont pas homogènes}$$

Exemple

- Effectifs observés :

Partic. / Ech	collège A	collège B	$N_{k.}$
oui	12	26	38
non	38	34	72
$N_{.l}$	50	60	$n = 110$

- Effectifs théoriques :

Partic. / Ech	collège A	collège B
oui	17,27	20,73
non	32,73	39,27

- $(J_n)^{obs} = \frac{(12-17,27)^2}{17,27} + \ldots + \frac{(34-39,27)^2}{39,27} = 4,504 > x_{1,0.95} = 3.84$