Le test du χ^2

Jean-François Bergez et Jean-François Delmas

October 10, 2017

Contents

1 A-t-on autant de chance de naître pour chacun des jours de la semaine?

Un hôpital américain veut savoir si les naissances sont réparties équitablement sur les jours de la semaine. Il dispose de l'observation des naissances de l'année 1997, (source : NVSR 1999)

Jour	L	M	M	J	V	S	D
Naissances	564772	629408	609596	604812	605280	450840	404456

Table 1: Répartition des naissances suivant les jours de la semaine

Rappel: un peu de théorie

On suppose qu'il s'agit de la réalisation des n variables aléatoires indépendantes et de même loi $X_1, \ldots, X_n$ à valeurs dans $\{1, \ldots, m\}$. Ici $m = 7$ puisque les variables aléatoires prennent leurs valeurs dans l'ensemble des jours de la semaine. On cherche donc à connaître $p = (p_1, \ldots, p_m)$ où $p_i = \mathbb{P}(X_1 = i)$. Plus exactement on désire savoir si la loi $p = (p_1, \ldots, p_m)$ est égale à une loi donnée $p^0 = (p_1^0, \ldots, p_m^0)$. Dans l'exemple ci-dessus, on désire savoir si les naissances sont équidistribuées sur les jours de la semaine, soit $p_0 = (\frac{1}{7}, \ldots, \frac{1}{7})$, la loi uniforme

sur $\{1, \ldots, 7\}$. On veut donc savoir si l'hypothèse *les naissances sont équidistribuées* dite hypothèse nulle, notée $H_0 = \{p = p^0\}$, est réaliste ou non. On utilise pour cela le test d'adéquation du χ^2.

On définit la statistique

$$\zeta_n = n \sum_{i=1}^{m} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0}$$

où $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)$ est le vecteur des **fréquences empiriques** :

$$\hat{p}_i = \frac{N_i}{n},$$

N_i étant le nombre d'occurrences de i dans l'échantillon de taille n.

Si H_0 est vrai, alors ζ_n converge en loi vers un χ^2 à $m-1$ degré de liberté. En particulier ζ_n prend les valeurs typiques du χ^2 à $m-1$ degré de liberté. En revanche si H_0 n'est pas vraie i.e. $p \neq p_0$, alors $\zeta_n \to +\infty$ quand $n \to +\infty$. Ainsi ζ_n prend de grandes valeurs. On rejette donc H_0 si on observe des valeurs anormalement grandes pour ζ_n, par exemple supérieures à un z donné. Toutefois, comme la loi du χ^2 est portée par $\mathbb{R}^+$, toutes les valeurs de $[0, +\infty[$ sont possibles. Il existe donc une probabilité α de rejeter à tort H_0 (risque de première espèce). La valeur de α dépend de z:

$$\mathbb{P}(\zeta_n \geqslant z) = \mathbb{P}(\chi^2(m-1) \geqslant z) = \alpha$$

La pratique

Utilisation de ce résultat.

- On dispose de l'observation de l'échantillon $X_1 = x_1, \ldots, X_n = x_n$.
- On dispose de la loi $p^0 = (p_1^0, \ldots, p_m^0)$ présumée des variables aléatoires (indépendantes) $X_1, \ldots, X_n$.
- On calcule les occurrences N_i de l'entier i à partir des x_i.
- On calcule la valeur de $\zeta_n = z_n$.
- On calcule $\boxed{\alpha_n = \mathbb{P}(\chi^2(m-1) \geqslant z_n)}$.
- Si α_n prend des valeurs faibles, inférieures, à $\alpha = 5\%$ typiquement, alors on rejette l'hypothèse H_0 sinon on l'accepte (voir la figure (1))
- α représente le niveau de confiance du test.

Ce risque α incontournable dépend du contexte et est fixé par le décisionnaire. Traditionnellement $\alpha = 5\%$, mais on peut choisir des valeurs bien plus faibles pour des domaines sensibles.

La mise en oeuvre

On revient au problème initial. On dispose pour cela des nombres des naissances présentés dans le tableau (1).

```
// les occurences des naissances
N=[564772,629408,609596,604812,605280,450840,404456];
// la loi uniforme sur {1,...,7}
p0=ones(1,7)/7;
```

La fonction `test_chi2(N,p0)` retourne la p-valeur du test du χ^2 d'adéquation de loi: `N` est le vecteur ligne des occurrences observées, et `p0` est le vecteur ligne des probabilités d'occurrence sous l'hypothèse nulle. Cliquer sur le lien ci-dessus pour obtenir le code de la fonction, le copier dans la fenêtre scilab ou le sauvegarder, sous le nom `test_chi2.sce`, dans le répertoire où vous utilisez scilab. Dans ce dernier cas, pour charger la fonction, utiliser la commande: `getf 'test_chi2.sce'`.

Pour appliquer la fonction, utiliser la commande:

```
//execution du test
alpha=test_chi2(N,p0)
```

Question 1 *les naissances se répartissent-elles équitablement sur les jours de la semaine?*

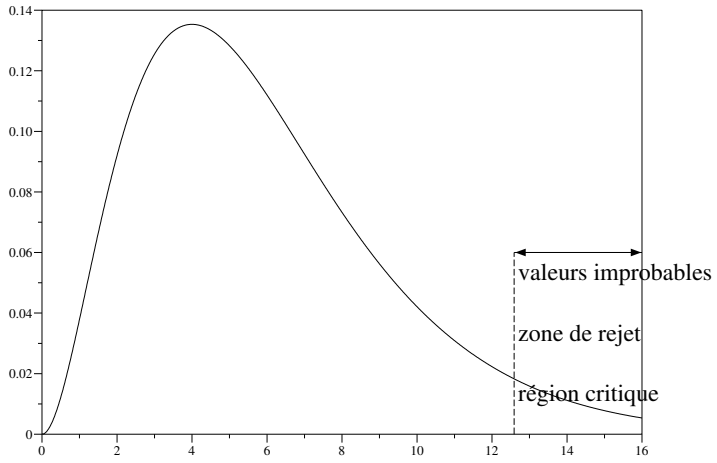

Figure 1: Densité de $\chi^2(6)$

2 Test d'un générateur de nombres aléatoires

On veut tester la validité d'un générateur de nombres aléatoires à valeurs dans $\{0,\dots,m\}$ et de loi uniforme. On génère n nombres suivant cette loi. On applique le test du χ^2 et on rejette à 5%. On effectue cette opération K fois et on compte le nombre de rejets. En théorie, on devrait obtenir en première approximation $0.05 \times K$ rejets puisque on simule suivant une loi uniforme et que l'on rejette à 5%. On génère un vecteur ligne de n réalisations d'une loi uniforme sur $\{0,\dots,9\}$, ici $m = 9$:

```
X=grand(1,n,'uin',0,m)
```

Ensuite la fonction occurrence retourne `N` le nombre d'occurrences des chiffres $0,\dots,9$ dans X (cliquer sur le lien ci-dessus pour obtenir le code de la fonction et le recopier dans la fenêtre scilab)

La fonction `test_generateur(n,K,m)` retourne le nombre de rejets du test d'adéquation du χ^2 pour `K` tests effectués sur des échantillons de `n` données du générateur de scilab de la loi uniforme sur $\{0,\dots,$`m`$\}$.

On obtient par exemple le résultat suivant :

K	n	Nombre de rejets obtenus
1000	100	53

Question 2 *Quelle est la loi du nombre de rejets? Donner, pour n et K grands, un intervalle auquel appartient le nombre de rejets avec probabilité de 99% quand le générateur de nombres aléatoires est parfait.*

On réalise la même expérience mais avec une loi p_1 un peu différente de la loi p_0 uniforme sur $\{0,\dots,9\}$. Remarquons que la loi des occurrences lors de n simulations suivant la loi p_1 est exactement la loi multinômiale de paramètre (n, p_1). Cette dernière loi est directement simulable à l'aide de

```
grand.
```

On choisit une probabilité p_1 qui diffère de p_0 seulement pour les probabilités d'apparition de 8 et 9.

```
// loi uniforme sur {0,...,m}
p0=ones(1,m+1)/(m+1);
p1=p0;
// $ designe la derniere coordonnee du vecteur
p1($)=0.15;
p1($-1)=0.05;

// on genere une variable de loi multinomiale de parametre (n,p1)
grand(1,'mul',n,p1([1:$-1])')
```

On simule K fois suivant cette loi p_1 et on compte le nombre de rejets. Pour cela on utilise la fonction suivante: `test_generateur_faux(n,K,m,p1)`, qui retourne le nombre de rejets du test d'adéquation du χ^2 pour `K` tests effectués sur des échantillons de `n` données du générateur de scilab de la loi `p1`.

On obtient alors le résultat suivant pour une simulation particulière :

K	n	Nombre de rejets obtenus
1000	100	281

Le nombre de rejets est plus important et à juste titre puisqu'on n'a pas simulé selon une loi uniforme. Cependant on ne rejette pas systématiquement car la loi simulée "ressemble" à une loi uniforme.

Cette fois-ci on remplace la loi uniforme par une loi binômiale $\mathcal{B}(9, 1/2)$:

```
p1=binomial(1/2,9);
```

Question 3 *Simuler le nombre de rejets pour $K = 1000$ simulations et $n = 100$. Que remarquez vous?*

Les histogrammes ci-dessous permettent de visualiser la différence entre la loi uniforme sur $\{0, \ldots, 9\}$ et la loi binômiale $\mathcal{B}(9, 1/2)$.

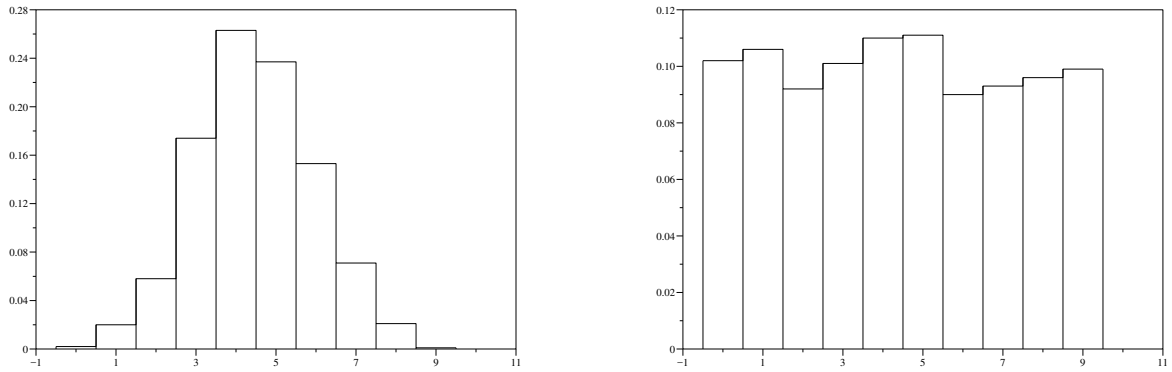

Figure 2: Simulation de 1000 variables de loi binômiale $\mathcal{B}(9, 1/2)$ et loi uniforme sur $\{0, \ldots, 9\}$

3 Les dés de Weldon

Weldon a réalisé $n = 26306$ lancers de 12 dés à 6 faces. On note X_i le nombre de faces comportant un cinq ou un six lors du i-ème lancer. Ces variables aléatoires prennent leurs valeurs dans $\{0, \ldots, 12\}$. Les fréquences empiriques observées sont $f_j = \frac{N_j}{n}$ où N_j est le nombre de lancers où l'événement "*on a observé j faces comportant un* 5 *ou un* 6" se realise : $N_j = \sum_{i=1}^{n} 1_{\{X_i=j\}}$. Les résultats sont les suivants : (source : FELLER Tome 1 pp. 148-149 *An introduction to probability theory and its applications* (1968))

f_0	f_1	f_2	f_3	f_4	f_5	f_6
0.007033	0.043678	0.124116	0.208127	0.232418	0.197445	0.116589

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
0.050597	0.015320	0.003991	0.000532	0.000152	0.000000

Quand les dés ne sont pas biaisés la probabilité d'observer les faces 5 ou 6 dans un lancer est 1/3. Les X_i suivent donc une loi binômiale $p_0 = \mathcal{B}(12, 1/3)$. Le vecteur N représente les occurrences correspondant aux fréquences empiriques.

```
N=[185,1149,3265,5475,6114,5194,3067,1331,403,105,14,4,0];
p0=binomial(1/3,12);
```

Question 4 *Les dés sont-ils biaisés?*

Question 5 *Dans les cas où les dés sont biaisés, avec tous le même biais, estimer la probabilité d'obtenir un* 5 *ou un* 6 *lors d'un lancer. Vérifier en utilisant un test du* χ^2*, dont on justifiera le nombre de degrés de liberté, que les données proviennent bien de la loi binômiale* $\mathcal{B}(12, p)$*, où* p *est inconnu. (Attention, il faut changer le code de la fonction* `test_chi2(N,p0)`*).*

Question 6 (Facultatif) *Le fait que les données proviennent d'une loi binômiale* $\mathcal{B}(12, p)$ *implique-t-il que les 12 dés ont même biais? Quelle est donc la conclusion à la question précédente?*

4 Un autre problème de naissance

On désire étudier la répartition des naissances suivant le type du jour de semaine (jours ouvrables ou week-end) et suivant le mode d'accouchement (naturel ou par césarienne). Les données proviennent du "National Vital Statistics Report" et concernent les naissances aux USA en 1997.

Naissances	Naturelles	César.	Total
J.O.	2331536	663540	2995076
W.E.	715085	135493	850578
Total	3046621	799033	3845654

Naissances	Naturelles	César.	Total
J.O.	60.6 %	17.3 %	77.9%
W.E.	18.6 %	3.5 %	22.1%
Total	79.2 %	20.8 %	100.0%

```
N=[2331536 715085 663540 135493];
```

On note $p_{J,N}$ la probabilité qu'un bébé naisse un jour ouvrable et sans césarienne, $p_{W,N}$ la probabilité qu'un bébé naisse un week-end et sans césarienne, $p_{J,C}$ la probabilité qu'un bébé naisse un jour ouvrable et par césarienne, $p_{W,C}$ la probabilité qu'un bébé naisse un week-end et par césarienne.

Question 7 *Donner l'estimateur des fréquences empiriques de* $p = (p_{J,N}, p_{W,N}, p_{J,C}, p_{W,C})$.

Question 8 *À l'aide d'un test du* χ^2*, pouvez-vous accepter ou rejeter l'hypothèse d'indépendance entre le type du jour de naissance (jour ouvrable ou week-end) et le mode d'accouchement (naturel ou césarienne)?*

Question 9 *On désire savoir s'il existe une évolution significative dans la répartition des naissances par rapport à 1996. À l'aide d'un test du* χ^2*, pouvez-vous accepter ou rejeter l'hypothèse* $p = p_0$*, où* p_0 *correspond aux données de 1996? On donne les valeurs suivantes pour* p_0*:*

Naissances	Naturelles	Césariennes
J.O.	60.5 %	17.0 %
W.E.	18.9 %	3.6 %

```
p0=[60.5 18.9 17 3.6]/100;
```