

SCILAB à l'École nationale des ponts et chaussées

<http://www.enpc.fr/scilab>

---

Statistique  
*Le test du  $\chi^2$*

---

Jean-François BERGEZ

Jean-François DELMAS

dernière date de mise à jour : 23 avril 2003

## Table des matières

|   |   |   |
|---|---|---|
| 1 | A-t-on autant de chance de naître pour chacun des jours de la semaine ? | 1 |
| 2 | Test d'un générateur de nombres aléatoires                              | 4 |
| 3 | Puissance d'un test (FACULTATIF)  | 7 |
| 4 | Les dés de Weldon   | 8 |
| 5 | Un autre problème de naissance  | 9 |

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

Un hôpital américain veut attribuer les congés de ses obstétriciens pour l'année 1998. 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 |

TAB. 1 – Répartition des naissances suivant les jours de la semaine

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

## Un peu de théorie

On suppose qu'il s'agit de la réalisation des  $n$  variables aléatoires indépendantes et de même loi  $X_1, \dots, X_n$  à valeurs dans  $\{1, \dots, m\}$ . Ici  $m = 7$  puisque les variables aléatoires prennent leurs valeurs dans l'ensemble des jours de la semaine. On cherche donc à connaître  $p = (p_1, \dots, p_m)$  où  $p_i = \mathbb{P}(X_1 = i)$ . Plus exactement on désire savoir si la loi  $p = (p_1, \dots, p_m)$  est égale à une loi donnée  $p^0 = (p_1^0, \dots, p_m^0)$ . Dans l'exemple ci-dessus, on désire savoir si les naissances sont équidistribuées sur les jours de la semaine, soit  $p_0 = (\frac{1}{7}, \dots, \frac{1}{7})$ , la loi uniforme sur  $\{1, \dots, 7\}$ . 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  $\chi^2$ .

On définit la statistique

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

où  $\hat{p} = (\hat{p}_1, \dots, \hat{p}_m)$  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  $\zeta_n$  converge en loi vers un  $\chi^2$  à  $m - 1$  degré de liberté. En particulier  $\zeta_n$  prend les valeurs typiques du  $\chi^2$  à  $m - 1$  degré de liberté. En revanche si  $H_0$  n'est pas vraie i.e.  $p \neq p_0$ , alors  $\zeta_n \rightarrow +\infty$  quand  $n \rightarrow +\infty$ . Ainsi  $\zeta_n$  prend de grandes valeurs. On rejette donc  $H_0$  si on observe des valeurs anormalement grandes pour  $\zeta_n$ , par exemple supérieures à un  $z$  donné. Toutefois, comme la loi du  $\chi^2$  est portée par  $\mathbb{R}^+$ , toutes les valeurs de  $[0, +\infty[$  sont possibles. Il existe donc une probabilité  $\alpha$  de rejeter à tort  $H_0$  (risque de première espèce). La valeur de  $\alpha$  dépend de  $z$  :

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

## La pratique

Utilisation de ce résultat.

- On dispose de l'observation de l'échantillon  $X_1 = x_1, \dots, X_n = x_n$ .
- On dispose de la loi  $p^0 = (p_1^0, \dots, p_m^0)$  présumée des variables aléatoires (indépendantes)  $X_1, \dots, X_n$ .
- 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  $\alpha_n = \mathbb{P}(\chi^2(m-1) \geq z_n)$ .
- Si  $\alpha_n$  prend des valeurs faibles, inférieures, à  $\alpha = 5\%$  typiquement, alors on rejette l'hypothèse  $H_0$  sinon on l'accepte.
- $\alpha$  représente le niveau de confiance du test.

Ce risque  $\alpha$  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 œuvre

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

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

On définit la fonction qui calcule la p-valeur :

```
// cette fonction renvoie P(chi2>zeta_n)

// N est une matrice des effectifs observés
// p0 est une matrice dont chaque ligne est une probabilité

// chaque ligne de N est le vecteur des occurrences et la ligne
// correspondante de p0 est la loi par rapport à laquelle on teste.

function[proba]=chi2(N,p0)
    n=sum(N,'c');// cardinal de l'échantillon observé
    // on vérifie que tous les échantillons ont même taille.
    if norm(n(1)*ones(n)- n) > %eps then
error("échantillons de tailles différentes");
return
    end
    n = n(1);
    // calcul de zeta_n
    zeta_n=n*sum(((N/n-p0).^2)./p0,'c');
    // nombre de degrés de liberté (= nombre de classes dans N-1)
    d= size(N,'c')-1
    // on calcule la proba pour un chi 2 à d-1 degrés d'être supérieur à zeta
    [p,q]=cdfchi("PQ",zeta_n,d*ones(zeta_n));
    proba=q;
endfunction;
```

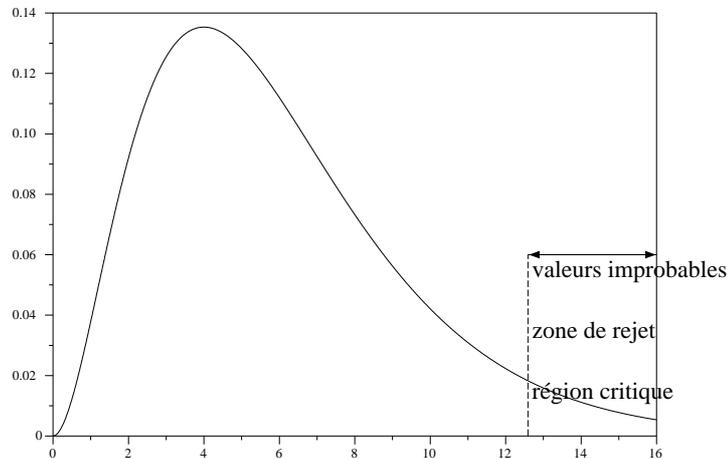
On calcule ainsi  $\zeta_n = 85210$  et  $\alpha_n = \mathbb{P}(\chi^2(6) > 85210) = 0$ . Comme on sait que  $\mathbb{P}(\chi^2(6) > 23) = 0.001$  on rejette cette hypothèse au niveau 0.1%. Les naissances ne se répartissent pas de manière uniforme sur les jours de la semaine.

```
//exécution du test
alpha=chi2(N,p0);
```

## Mode d'emploi

– Taper ce qui précède dans un fichier script “prog1.sce”.

- Exécuter dans la fenêtre Scilab en tapant la commande  
`exec prog1.sce;`
- Pour récupérer le résultat tapez dans la fenêtre :  
`alpha`



GRAPH. 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  $\chi^2$  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 on compte le nombre d'occurrences des chiffres  $0, \dots, 9$  dans  $X$  :

```
function[N]=occurrences(X,m)
// X est une matrice dont chaque ligne est la réalisation
// de variables aléatoires à valeurs
// dans {0,...,m}
mx= size(X,'r');
// taille de l'échantillon
N=zeros(mx,m+1);
for i=1:m+1
N(:,i)=sum(X==i-1,'c');
end
endfunction;
```

Puis on utilise la fonction

```
// on teste K fois le générateur avec n tirages de la loi uniforme (sur
// {0,...,m} à chaque test.
```

```
function[nb_rejets]=test_generateur(n,K,m)
// K fois n réalisations de la loi uniforme sur {0,...,m}
X=grand(K,n,'uin',0,m);
// calcul des nombres d'occurrences de 1,...,m dans X
No=occurrences(X,m);
// calcul des p-valeurs
alpha=chi2(No,ones(K,m+1)/(m+1));
// nombre de rejets à 5%
nb_rejets = sum(alpha<=0.05);
endfunction;
```

et on obtient pour une simulation :

| $K$  | $n$ | Nombre de rejets obtenus |
|------|-----|--------------------------|
| 1000 | 100 | 53                       |

**Question 2.** *Quelle est la loi asymptotique (quand  $K \rightarrow \infty$ ) du pourcentage 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;
// $ désigne la dernière coordonnée du vecteur
p1($)=0.15;
p1($-1)=0.05;
```

```
////////////////////////////////////
//
// ON ÉLIMINE LA DERNIÈRE COORDONNÉE DE P1 POUR
// UTILISER LE GÉNÉRATEUR grand('mul')
//
////////////////////////////////////
```

```
p1_tilde=p1([1:$-1]);
// on génère une variable de loi multinômiale de paramètre (n,p1)
grand(1,'mul',n,p1_tilde')
```

On simule  $K$  fois suivant cette loi  $p_1$  et on compte le nombre de rejets. Pour cela on utilise la forme vectorielle suivante :

```
// simule K fois les occurrences de n simulations selon p1
// (p1_tilde se déduit de p1 en supprimant la dernière valeur de p1)
X=grand(K,'mul',n,p1_tilde)';
// on calcule les p-valeurs pour les K simulations
r=chi2(X,ones(K,1)*p0);
// on détermine le nombre de rejets
nb_rejets= sum(r<=0.05)
```

On obtient alors le résultat suivant pour une simulation particulière :

| $K$  | $n$ | Nombre de rejets obtenus |
|------|-----|--------------------------|
| 1000 | 100 | 281                      |

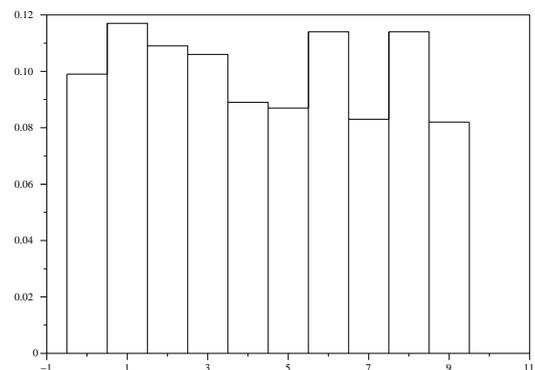
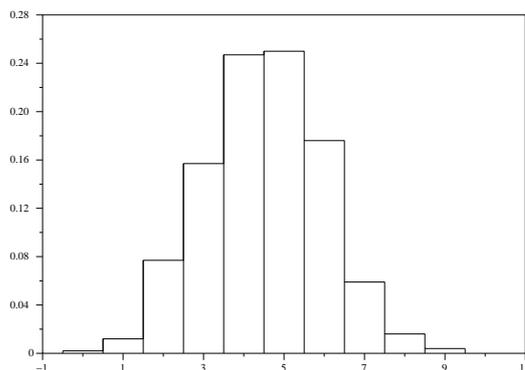
Le nombre de rejets est plus important et à juste titre puisqu'on n'a pas simulé selon une loi uniforme. Cependant on ne rejette pas systématiquement car la loi simulée "ressemble" à une loi uniforme.

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  $\{0, \dots, 9\}$  et la loi binômiale  $\mathcal{B}(9, 1/2)$ .



GRAPH. 2 – Simulation de 1000 variables de loi binômiale  $\mathcal{B}(9, 1/2)$  et loi uniforme sur  $\{0, \dots, 9\}$

### 3 Puissance d'un test (FACULTATIF)

On détermine la puissance du test du  $\chi^2$  sur l'exemple suivant : on simule  $n$  tirages suivant une loi  $p_1$  de la forme :

$$p_1 = \left( \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10} - d, \frac{1}{10} + d \right).$$

On calcule alors la valeur de  $\zeta_n$  puis on rejette  $H_0$  à 5%. On effectue  $K$  fois cette opération et on en déduit une estimation de la probabilité  $r_{p_1} = \frac{\text{nombre de rejets}}{K}$  de rejeter  $H_0$  quand on simule  $p_1$  : c'est la puissance du test. Enfin on trace  $r_{p_1}$  en fonction de la pseudo-distance du  $\chi^2$  entre  $p_0$  et  $p_1$  :

$$D(p_1, p_0) = \sum_{i=0}^9 \frac{(p_1(i) - p_0(i))^2}{p_0(i)}.$$

Elle est calculée grâce à la fonction

```
//pseudo distance du chi2
//p1 est la loi à tester, p0 est la loi de référence
fonction[d]=Dist(p1,p0)
d=sum((p1-p0).^2)./p0;
endfonction;
```

Le programme qui réalise ceci est similaire à celui du test des générateurs :

```
// On effectue K fois n simulations suivant la loi p1, et on obtient la
// probabilité de rejet comme fonction de la distance
fonction [nb_rejets,D]=test(n,K,m,d)
    p0=ones(1,m+1)/(m+1); // loi uniforme sur {0,...,m}
    p1=p0;
    // avant de modifier p0, on vérifie que d est une valeur possible
    if max(p1($)-1,-p1($-1)) >= d | d >= min(1-p1($-1),p1($)) then
error("d est incompatible")
    end
    p1($)= p1($) -d;
    p1($-1)= p1($-1) + d;
    // calcul de la pseudo distance du chi2
    // entre p0 et p1
    D=Dist(p1,p0);
    p1_tilde=p1([1:$-1]);
    X=grand(K,'mul',n,p1_tilde)';
    // simule K fois les occurrences de n simulations selon p1.
    r=chi2(X,ones(K,1)*p0);
    nb_rejets= sum(r<=0.05);
endfonction;
```

Les résultats sont récapitulés dans le tableau suivant pour  $d \in \{0, 0.01, \dots, 0.09\}$ .

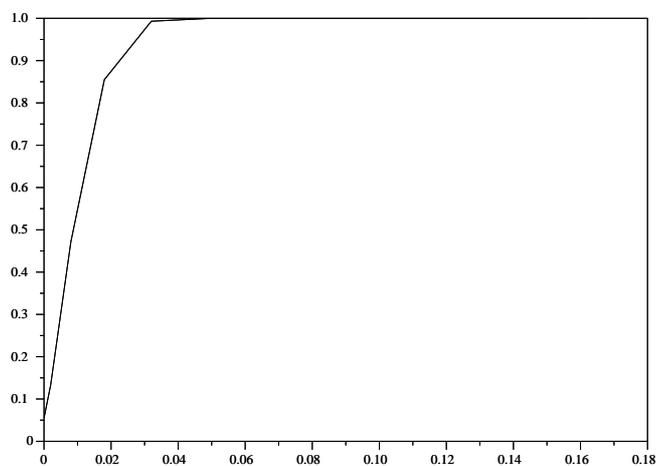
```
//pour la courbe de puissance du test
function[D,r]=df(k)
    n = 1000;
    K = 1000;
    m = 9;
    r = ones(1,k);
    D = ones(1,k);
    for i=1:k
d =(i-1)/k/(m+1)
[r1,D1]=test(n,K,m,d);
r(i)=r1/K;
D(i)=D1;
    end
    plot2d(D,r);
endfunction;
```

|               |       |       |       |       |       |       |       |       |       |       |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $D(p_1, p_0)$ | 0.000 | 0.002 | 0.008 | 0.018 | 0.032 | 0.050 | 0.072 | 0.098 | 0.128 | 0.162 |
| $r_{p_1}$     | 0.042 | 0.127 | 0.455 | 0.873 | 0.993 | 1     | 1     | 1     | 1     | 1     |

La courbe est obtenue en tapant :

```
df(10);
```

```
;
```



GRAPH. 3 – Puissance du test du  $\chi^2$

## 4 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

$\{0, \dots, 12\}$ . 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 réalise :  $N_j = \sum_{i=1}^n 1_{\{X_i=j\}}$ . 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 binomiale  $p_0 = \mathcal{B}(12, 1/3)$ . Le vecteur  $N$  représente les occurrences correspondant aux fréquences empiriques.

$N=[185, 1149, 3265, 5475, 6114, 5194, 3067, 1331, 403, 105, 14, 4, 0]$  ;  
 $p_0=\text{binomial}(1/3, 12)$  ;

**Question 4.** *Les dés sont-ils biaisés ?*

**Question 5.** *Dans les cas où les dés sont biaisés, avec tous le même biais, calculer la probabilité d'obtenir un 5 ou un 6 lors d'un lancer. Vérifier en utilisant un test du  $\chi^2$ , dont on justifiera le nombre de degrés de liberté, que les 12 dés ont le même biais, i.e. les données proviennent bien de la loi binomiale  $\mathcal{B}(12, p)$ , où  $p$  est inconnu.*

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

**Question 7.** À l'aide d'un test du  $\chi^2$ , 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 8.** 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  $\chi^2$ , 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 %       |

$p_0 = [60.5 \ 18.9 \ 17 \ 3.6] / 100;$