TESTS DU KHI-DEUX

Aimé Lachal 1999/2000

Ajustement à une loi de Bernoulli

Expérience : dans une population de 100 000 nouveau-nés, on recense 51 400 garçons et 48 600 filles.

- *Variables aléatoires :*
  - caractère X : sexe du nourrisson ; X suit une loi de Bernoulli de paramètre $p_1 = \mathbb{P}(X = \text{sexe masculin})$; $p_2 = 1 - p_1$;
  - échantillon $(X_1, \ldots, X_n)$, $n = 100\,000$;
  - X_k : sexe du k^{e} nourrisson recensé ;
  - I_1 : classe de sexe masculin ;
  - I_2 : classe de sexe féminin ;
  - $N_1 = \text{card}\{k \in \{1, \ldots, n\} : X_k \in I_1\}$: nombre de garçons ;
  - $N_2 = \text{card}\{k \in \{1, \ldots, n\} : X_k \in I_2\}$: nombre de filles.

Hypothèse H_0 : $p_1 = \frac{1}{2}$.

Variable de décision : $Z_n = \chi^2 \equiv \displaystyle\sum_{i=1}^{2} \frac{(N_i - np_i)^2}{np_i} = \frac{(N_1 - np_1)^2}{np_1(1-p_1)}$.

$r = 2 \to 2 - 1 = 1$ d.d.l. et alors : $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(1)$.

- *Risque :* $\alpha = \mathbb{P}(\chi^2(1) > k_{1-\alpha}(1)) = \mathbb{P}(\text{rejet de } H_0 / H_0 \text{ vraie})$.

$$\chi^2_{\text{observé}} = \frac{(51\,400 - 50\,000)^2}{50\,000} + \frac{(48\,600 - 50\,000)^2}{50\,000} = \frac{(51\,400 - 50\,000)^2}{25\,000} = 78{,}4$$

Pour $\alpha = 0{,}01$, on a $\chi^2_{\text{observé}} > 6{,}64 = k_{0,99}(1)$
$\Longrightarrow$ **on rejette H_0 au risque de 1/100.**

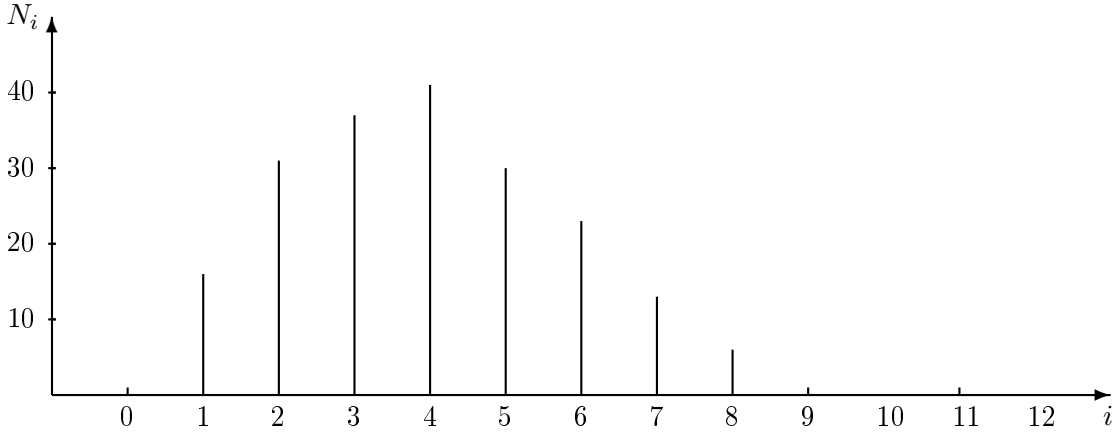
- *Autres données :* 50 400 garçons et 49 600 filles.
  $\chi^2_{\text{observé}} = 6{,}4 < 6{,}64 = k_{0,99}(1) \Longrightarrow$ **on accepte H_0.**

Ajustement à une loi de Poisson

Expérience : contrôle de qualité lors de la fabrication en grande série d'une pièce. On compte le nombre de pièces défectueuses dans une série de 400 pièces. On opère ainsi sur 200 séries, soit 800 000 pièces.

- *Variables aléatoires :*
  - caractère X : nombre de pièces défectueuses par série ;
  - échantillon $(X_1, \ldots, X_n)$, $n = 200$;
  - X_k : nombre de pièces défectueuses dans la k^{e} série ;
  - $I_i = \{i\}$, $0 \leqslant i \leqslant 400$;
  - $N_i = \text{card}\{k \in \{1, \ldots, n\} : X_k \in I_i\}$: nombre de séries donnant i pièces défectueuses.
- *Données :* série statistique $\{(i, N_i), 0 \leqslant i \leqslant 400\}$:

i	0	1	2	3	4	5	6	7	8	9	10	11	$\geqslant 12$
N_i	1	16	31	37	41	30	23	13	6	1	0	1	0
np_i	3,66	14,65	29,31	39,07	39,07	31,26	20,84	11,91	5,95	2,65	1,06	0,38	0,18
	18,31								10,22				



$\longrightarrow$ Ajustement à une loi de Poisson de paramètre estimé $\lambda_{\text{observé}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{400} iN_{i,\text{observé}} = 4$.

Hypothèse H_0 : $\forall i \in \mathbb{N},\ p_i = \mathbb{P}(X = i) = e^{-4}\dfrac{4^i}{i!}$.

On modifie les classes pour que $\forall i \in \mathbb{N},\ np_i \geqslant 5$.

Variable de décision : $Z_n = \chi^2 \equiv \displaystyle\sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$.

$r = 8 \to 8 - 1 - 1 = 6$ d.d.l. et alors : $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(6)$.

- *Risque :* $\alpha = \mathbb{P}(\chi^2(1) > k_{1-\alpha}(1)) = \mathbb{P}(\text{rejet de } H_0 / H_0 \text{ vraie})$.

  Pour $\alpha = 0{,}05$ on a $\chi^2_{\text{observé}} = 1{,}29 < 12{,}59 = k_{0,95}(1) \Longrightarrow$ **on accepte H_0.**

Ajustement à une loi de Gauss

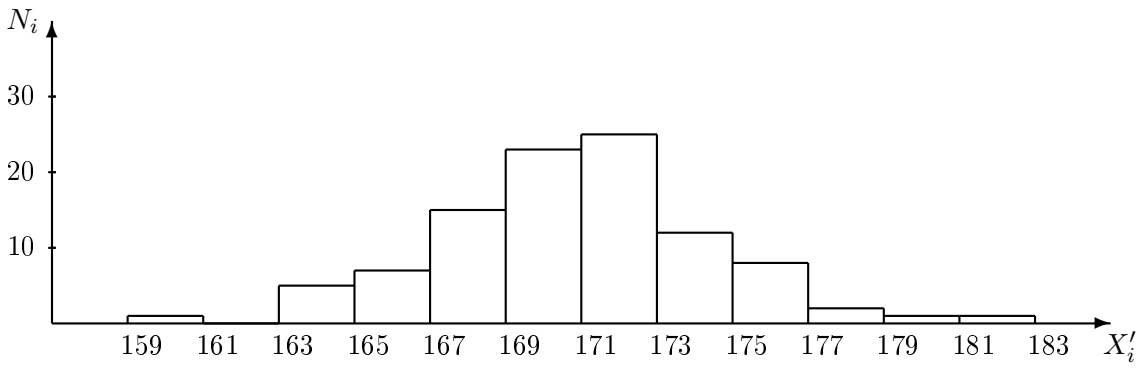
Expérience : statistique des tailles des individus d'une population d'effectif 100.

• *Variables aléatoires :*
- caractère X : taille en cm ;
- échantillon $(X_1, \ldots, X_n)$, $n = 100$;
- X_k : taille du k^{e} individu ;
- $I_i = [159 + 2i; 161 + 2i]$, $0 \leqslant i \leqslant 11$;
- $N_i = \text{card}\{k \in \{1, \ldots, n\} : X_k \in I_i\}$: nombre d'individus de taille comprise entre $159 + 2i$ cm et $161 + 2i$ cm.

• *Données :* série statistique $\{(X'_i, N_i), 0 \leqslant i \leqslant 11\}$ où $X'_i = 160 + 2i$ est le centre de la classe I_i :

X'_i	160	162	164	166	168	170	172	174	176	178	180	182
N_i	1	0	5	7	15	23	25	12	8	2	1	1

• *Histogramme :*



$\longrightarrow$ Ajustement à une loi de Gauss de paramètres estimés

$$m_{\text{observé}} = \frac{1}{n}\sum_{i=0}^{11} N_i X'_i = 170{,}86 \quad \text{et} \quad \sigma^2_{\text{observé}} = \frac{1}{n}\sum_{i=0}^{11} N_i (X'_i - m)^2 = 13{,}30.$$

Hypothèse H_0 **:** $\forall i \in \mathbb{N},\ p_i = \mathbb{P}(X \in I_i) = \displaystyle\int_{159+2i}^{161+2i} e^{-\frac{(x-m)^2}{2\sigma^2}} \frac{dx}{\sqrt{2\pi}\sigma}$

$$= \Phi\left(\frac{2i - 9{,}86}{3{,}65}\right) - \Phi\left(\frac{2i - 11{,}86}{3{,}65}\right)$$

avec $\Phi(z) = \displaystyle\int_{-\infty}^{z} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}}$.

On modifie les classes pour que $\forall i \in \mathbb{N},\ np_i \geqslant 5$:

i	0	1	2	3	4
$\frac{1}{\sigma}(I_i - m)$	$]-\infty;-2{,}70]$	$[-2{,}70;-2{,}15]$	$[-2{,}15;-1{,}60]$	$[-1{,}60;-1{,}06]$	$[-1{,}06;-0{,}51]$
N_i	1	0	5	7	15
np_i	0,35	1,23	3,90	8,98	16,04
	$\underbrace{\qquad\qquad\qquad}_{5{,}48}$				

5	6	7	8	9	10	11
$[-0{,}51;0{,}04]$	$[0{,}04;0{,}59]$	$[0{,}59;1{,}13]$	$[1{,}13;1{,}68]$	$[1{,}68;2{,}23]$	$[2{,}23;2{,}78]$	$[2{,}78;+\infty[$
23	25	12	8	2	1	1
21,1	20,64	14,84	8,27	3,36	1,02	0,27
			$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{12{,}92}$			

Variable de décision : $Z_n = \chi^2 \equiv \displaystyle\sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$.

$r = 7 \to 7 - 2 - 1 = 4$ d.d.l. et alors : $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(4)$.

- *Risque :* $\alpha = \mathbb{P}(\chi^2(1) > k_{1-\alpha}(4)) = \mathbb{P}(\text{rejet de } H_0 / H_0 \text{ vraie})$.

$$\chi^2_{\text{observé}} = \frac{(6-5{,}48)^2}{5{,}48} + \frac{(7-8{,}98)^2}{8{,}98} + \cdots + \frac{(12-14{,}84)^2}{14{,}84} + \frac{(12-12{,}92)^2}{12{,}92} = 2{,}21$$

Pour $\alpha = 0{,}05$, on a $\chi^2_{\text{observé}} < 7{,}78 = k_{0{,}95}(4) \Longrightarrow$ **on accepte** H_0**.**

Test d'indépendance

Expérience : dans une population de 60 000 individus, on étudie conjointement la couleur des cheveux (blonds, châtains ou bruns) et celles des sourcils (clairs ou foncés).

- *Variables aléatoires :*
  - caractère bivarié (X, Y) :
    X : couleur des sourcils ; X suit la loi de probabilité $\{p_{1.}, p_{2.}\}$:
    $p_{1.} = \mathbb{P}(X = \text{couleur claire})$ et $p_{2.} = \mathbb{P}(X = \text{couleur foncée})$;
    Y : couleur des cheveux ; Y suit la loi de probabilité $\{p_{.1}, p_{.2}, p_{.3}\}$:
    $p_{.1} = \mathbb{P}(Y = \text{couleur blonde})$, $p_{.2} = \mathbb{P}(Y = \text{couleur châtain})$
    et $p_{.3} = \mathbb{P}(Y = \text{couleur brune})$;
  - échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$, $n = 60\,000$;
  - X_k : couleur des sourcils du k^{e} individu ;
    Y_k : couleur des cheveux du k^{e} individu ;
  - caractères différenciés :
    X'_1 : couleur claire, X'_2 : couleur foncée ;
    Y'_1 : couleur blonde, Y'_2 : couleur châtain, Y'_3 : couleur brune.
  - $I_{1.}$ et $I_{2.}$: classes de couleurs claire et foncée ;
  - $I_{.1}$, $I_{.2}$ et $I_{.3}$: classes de couleurs blonde, châtain et brune ;
  - $N_{ij} = \text{card}\{k \in \{1, \ldots, n\} : (X_k, Y_k) \in I_{i.} \times I_{.j}\}$.
- *Données :* tableau de contingence $\{(X'_i, Y'_j, N_{ij}), 1 \leqslant i \leqslant 2, 1 \leqslant j \leqslant 3\}$:

sourcils X'_i \ cheveux Y'_j	blonds	châtains	bruns	totaux $N_{i.}$
clairs	9 468	2 105	3 364	14 937
foncés	3 238	30 472	11 353	45 063
totaux $N_{.j}$	12 706	32 577	14 717	60 000

Hypothèse H_0 : les v.a. X et Y sont indépendantes.

Effectifs théoriques : $np_{ij} = n\mathbb{P}((X, Y) \in I_{i.} \times I_{.j})$.

Sous H_0, $p_{ij} = p_{i.}p_{.j}$, et les effectifs théoriques estimés sont donnés par $\dfrac{N_{i.}N_{.j}}{n}$:

3 163,2	8 110,0	3 663,8	14 937
9 542,8	24 467,0	11 053,2	45 063
12 706	32 577	14 717	60 000

Variable de décision : $Z_n = \chi^2 \equiv \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{ij} - np_{ij})^2}{np_{ij}}$.

$r = 2$ et $s = 3 \to (2-1)(3-1) = 2$ d.d.l. et alors $Z_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(2)$

- *Risque :* $\alpha = \mathbb{P}(\chi^2(1) > k_{1-\alpha}(2)) = \mathbb{P}(\text{rejet de } H_0 / H_0 \text{ vraie})$.
  Pour $\alpha = 0{,}001$, $\chi^2_{\text{observé}} = 22\,684{,}9 > 13{,}82 = k_{0{,}999}(2)$
  $\Longrightarrow$ **on rejette H_0 au risque de 1/1 000.**