

Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

13 - Loi multinomiale et tableaux croisés

été 2022

Un peu de formalisme...

Tableau de comptage

X peut prendre les modalités $\{x_1, \dots, x_J\}$. On appelle **tableau de comptage** le vecteur \mathbf{n} de taille J $\mathbf{n} = [n_j] = (n_1, \dots, n_J)$ où n_j est le nombre d'individus dont la modalité est x_j .

Example Considérons l'exemple où X désigne la couleur des yeux, de la base `HairEyeColor`,

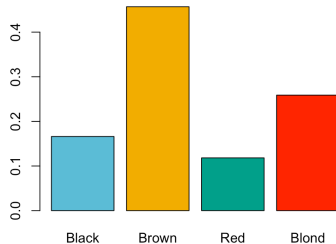
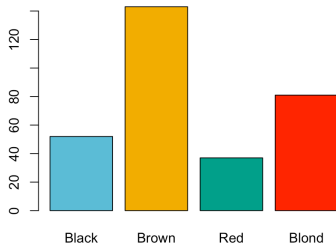
```
1 > data(HairEyeColor)
2 > n = apply(HairEyeColor[, , Sex="Female"], 1, sum)
3 > n
4 Black Brown Red Blond
5 52 143 37 81
```

Un peu de formalisme...

Si n est l'effectif total, $n = \sum_{i=1}^n \mathbf{1}_{j=x_i}$, et la fréquence est

$$\mathbf{f} = \frac{1}{n} \mathbf{n} = \left[\frac{n_j}{n} \right]$$

```
1 > barplot(n)
2 > f = n/sum(n)
3 > barplot(f)
```



avec le comptage (gauche) et les probabilités (droite)

Loi multinomiale

On suppose que $\{X_1, \dots, X_n\}$ est une collection de variables catégorielles indépendantes, de loi $\mathbf{p} = (p_1, \dots, p_J)$

La variable $Y_{j:i} = \mathbf{1}_j(X_i)$ suit une loi de Bernoulli $\mathcal{B}(p_j)$, où

$$p_j = \mathbb{E}[Y_j] = \mathbb{E}(\mathbf{1}_j(X)) = \mathbb{P}[X = x_j]$$

La variable $N_j = \sum_{i=1}^n \mathbf{1}_j(X_i) = \sum_{i=1}^n Y_{j:i}$ suit une loi binomiale $\mathcal{B}(n, p_j)$

Espérance, variance et covariance

$$\mathbb{E}(N_i) = np_i \quad \text{var}(N_i) = np_i(1 - p_i)$$

$$\text{cov}(N_i, N_j) = -np_i p_j$$

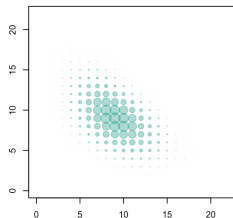
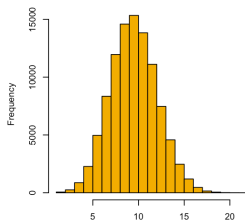
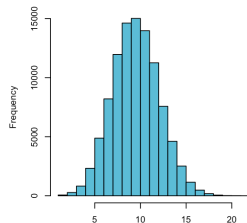
(admis)

Loi multinomiale

Les variables N_i et N_j ne sont pas indépendantes, car $\sum_{j=1}^J N_j = n$

Exemple On peut simuler une loi prenant les valeurs $\{1, 2, 3\}$, uniforme ($\mathbf{p} = (1/3, 1/3, 1/3)$), $n = 30$ fois.

```
1 > X = sample(1:3, size=n, replace=TRUE)
2 > N = table(X)[as.character(1:3)]
```



Loi multinomiale

On peut montrer que

Loi multinomiale

$$\mathbb{P}(N_1 = n_1, \dots, N_J = n_J) = \frac{n!}{n_1! \dots n_J!} p_1^{n_1} \dots p_J^{n_J}$$

pour tout $\mathbf{n} = (n_1, \dots, n_J)$ tel que $n_1 + \dots + n_J = n$.

En particulier si $J = 2$, on retrouve la loi binomiale,

$$\mathbb{P}(N_1 = n_1, N_2 = n_2) = \frac{n!}{n_1! n_2!} p_1^{n_1} p_2^{n_2}$$

pour n_1 et n_2 tels que $n_1 + n_2 = n$, ou

$$\mathbb{P}(N_1 = n_1, N_2 = n - n_1) = \frac{n!}{n_1! (n - n_1)!} p_1^{n_1} (1 - p_1)^{n - n_1}$$

Loi multinomiale

On peut montrer que

Loi multinomiale, approximation

Si $\{x_1, \dots, x_n\}$ est une collection de variables catégorielles indépendantes, de probabilités $\mathbf{p} = (p_1, \dots, p_J)$, et si n_j est le nombre d'observations de la modalité j ,

$$\frac{N_j - np_j}{\sqrt{np_j(1 - p_j)}} \approx \mathcal{N}(0, 1)$$

et

$$\sum_{j=1}^J \frac{(N_j - np_j)^2}{np_j} \approx \chi^2(J - 1)$$

(le résultat sera admis ici)

Cette dernière propriété permet de proposer un test de fréquence

Loi multinomiale, test $H_0 : \mathbf{p} = \mathbf{p}_0$ contre $H_1 : \mathbf{p} \neq \mathbf{p}_0$

Si $\{x_1, \dots, x_n\}$ est une collection de variables catégorielles indépendantes, de probabilités $\mathbf{p} = (p_1, \dots, p_J)$, pour tester $H_0 : \mathbf{p} = \mathbf{p}_0$ contre $H_1 : \mathbf{p} \neq \mathbf{p}_0$ la statistique de test est

$$Q = \sum_{j=1}^J \frac{(N\hat{p}_j - Np_{0,j})^2}{Np_{0,j}} = \sum_{j=1}^J \frac{(n_j - Np_{0,j})^2}{Np_{0,j}}$$

Si $H_0 : \mathbf{p} = \mathbf{p}_0$ est vraie, $Q \sim \chi^2(J-1)$. Et donc

► on rejette H_0 si $q > Q_{J-1}^{-1}(1 - \alpha)$

où Q_ν est la fonction de répartition de la loi du chi-deux, $\chi^2(\nu)$.

Loi multinomiale, test

On a lancé $n = 600$ fois un dé, est-il biaisé ?

```
1 > table(X1)
2      1      2      3      4      5      6
3    88    109    107    94    105    97
```

$$q = \frac{(88 - 100)^2}{100} + \frac{(109 - 100)^2}{100} + \frac{(107 - 100)^2}{100} + \frac{(94 - 100)^2}{100} + \frac{(105 - 100)^2}{100} + \frac{(97 - 100)^2}{100}$$

```
1 > sum((table(X1)-100)^2/100)
2 [1] 3.44
```

or le quantile à 95% d'une loi $\chi^2(6 - 1)$ est 11.07

```
1 > qchisq(.95,6-1)
2 [1] 11.0705
```

et la p -value vaut 36.7%

```
1 > 1-pchisq(3.44,6-1)
2 [1] 0.6324852
```

Loi multinomiale, test

On a lancé $n = 600$ fois un (autre) dé, est-il biaisé ?

```
1 > table(X2)
2      1      2      3      4      5      6
3     89    131     93     92    104     91
```

$$q = \frac{(89 - 100)^2}{100} + \frac{(131 - 100)^2}{100} + \frac{(93 - 100)^2}{100} + \frac{(92 - 100)^2}{100} + \frac{(104 - 100)^2}{100} + \frac{(91 - 100)^2}{100}$$

```
1 > sum((table(X2) - 100)^2 / 100)
2 [1] 12.92
```

qui dépasse le quantile à 95% d'une loi $\chi^2(6 - 1)$ est 11.07

```
1 > qchisq(.95, 6 - 1)
2 [1] 11.0705
```

et la p -value vaut 36.7%

```
1 > 1 - pchisq(12.92, 6 - 1)
2 [1] 0.0241401
```

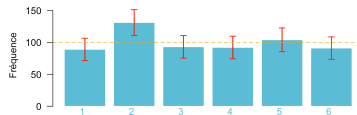
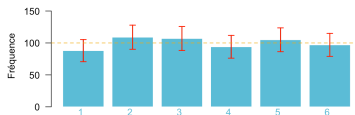
Tests multiples ★★★

On a ponctuellement des intervalles de confiance, sur les probabilités

$$\left[\hat{p}_j \pm u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}_j(1 - \hat{p}_j)}{n}} \right] \text{ où } \hat{p}_j = \frac{n_j}{n},$$

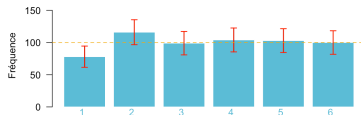
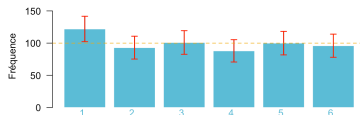
et sur les fréquences

$$\left[n_j \pm u_{1-\alpha} \cdot \sqrt{\frac{n_j(n - n_j)}{n}} \right]$$



Tests multiples ★★★

Ces intervalles de confiance sont associés à 6 tests simples

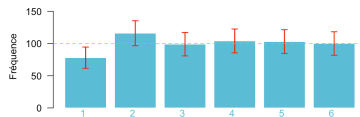
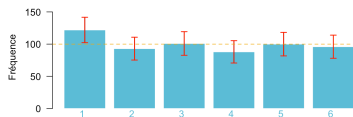


On peut refuser un test simple (un sur les six)

```
1 > table(X)
2   1   2   3   4   5   6
3  78 116  99 104 103 100
4 > prop.test(table(X)[1],600,1/6)
5
6 1-sample proportions test with continuity correction
7
8 data:  table(X)[1] out of 600, null probability 1/6
9 X-squared = 5.547, df = 1, p-value = 0.01851
10 alternative hypothesis: true p is not equal to
    0.1666667
11 95 percent confidence interval:
12  0.1046716 0.1601808
```

Tests multiples ★★★

et on peut accepter le test multiple (p -value de 17.6%)



Car ici, on regarde un test multiple, $H_0 : p_1 = \dots = p_6$.

```
1 > 1-pchisq(sum((table(X)-100)^2/100),6-1)
2 [1] 0.175996
```

i.e.

```
1 > chisq.test(table(X), p = rep(1/6,6))
2
3   Chi-squared test for given probabilities
4
5 data:  table(X)
6 X-squared = 7.66, df = 5, p-value = 0.176
```

Un peu de formalisme...

La formule de base repose sur

$$Z_j = \frac{\text{comptage observé} - \text{comptage attendu sous } H_0}{\sqrt{\text{comptage attendu}}} = \frac{O_j - E_j}{\sqrt{E_j}}$$

Si on a assez d'observations, $Z_j \approx \mathcal{N}(0, 1)$ et

$$Q = Z_1^2 + Z_2^2 + \dots + Z_{J-1}^2 + Z_J^2 \approx \chi^2(J-1)$$

que l'on notera aussi

$$Q = \sum_{j=1}^J \frac{(O_j - E_j)^2}{E_j} \approx \chi^2(J-1)$$

Un peu de formalisme...

De manière générale, la statistique de test est

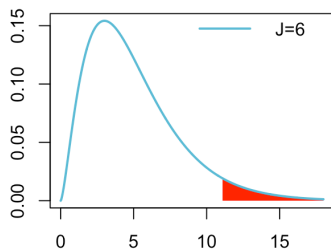
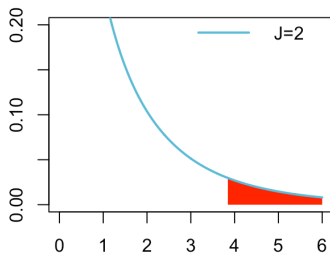
$$q = \sum_{j=1}^J \frac{(N\hat{p}_{0,j} - Np_{0,j})^2}{Np_{0,j}} = \sum_{j=1}^J \frac{(n_j - Np_{0,j})^2}{Np_{0,j}}$$

la p -value est

$$p = \mathbb{P}[Q > q | Q \sim \chi^2(J-1)]$$

mais on peut passer par la région de rejet

- ▶ si $q > Q_{J-1}^{-1}(1 - \alpha)$ on rejette H_0
- ▶ si $q < Q_{J-1}^{-1}(1 - \alpha)$ on ne rejette pas H_0



Test d'ajustement I

On a vu (partie 11) que le test de Komogorov Smirnov pouvait être utilisé comme test d'ajustement pour une loi continue. Pour des lois discrètes, on peut parfois utiliser un test du chi-deux.

Example Considérons la loi de Poisson $\mathcal{P}(2)$

```
1 > dpois(0:15,2),3
2 [1] 0.135 0.271 0.271 0.180 0.090 0.036 0.012 0.003
3 [9] 0.001 0.000 0.000 0.000 0.000 0.000 0.000 0.000
```

On pourra noter

```
1 > p = c(dpois(0:4,2),1-ppois(4,2))
2 > names(p) = c(0:4,"5+")
3 > p
4      0      1      2      3      4      5+
5 0.135 0.271 0.271 0.180 0.090 0.053
```


Test d'ajustement II

Test d'ajustement, test $H_0 : X \sim f_0$ contre $H_1 : X \not\sim f_0$

Si $\{x_1, \dots, x_n\}$ est une collection de variables indépendantes de loi f_0 . Notons $\mathbf{p}_0 = (p_{0,1}, \dots, p_{0,J})$ le vecteur de probabilités associées à f_0 , possiblement en faisant des regroupements de valeurs. Pour tester $H_0 : X \sim f_0$ contre $H_1 : X \not\sim f_0$ la statistique de test est

$$Q = \sum_{j=1}^J \frac{(N\hat{p}_j - Np_{0,j})^2}{Np_{0,j}} = \sum_{j=1}^J \frac{(n_j - Np_{0,j})^2}{Np_{0,j}}$$

Si $H_0 : \mathbf{p} = \mathbf{p}_0$ est vraie, $Q \sim \chi^2(J-1)$. Et donc

► on rejette H_0 si $q > Q_{J-1}^{-1}(1 - \alpha)$

où Q_ν est la fonction de répartition de la loi du chi-deux, $\chi^2(\nu)$.

Test d'ajustement III

Example 1 : Pendant la second guerre mondiale, les impacts de bombes V1 et V2 tombées dans une zone de 144 km^2 dans le sud de Londres. Il divisa cette zone en 576 zones de 0.25 km^2 et compta le nombre d'impact dans chacune des zones.

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226.74	229
1	211.39	211
2	98.54	93
3	30.62	35
4	7.14	7
5 and over	1.57	1
	576.00	576

On a le comptage complet

```
1 > (n=c(229,211,93,35,7,0,0,1))
2 [1] 229 211 93 35 7 0 0 1
```

Test d'ajustement IV

L'estimateur de la méthode des moments est $\hat{\lambda} = \bar{y}$

```
1 > (lambda = sum(n*((0:7))/sum(n)))
2 [1] 0.9322917
3 > y = rep(0:7, n)
4 > mean(y)
5 [1] 0.9322917
```

L'estimateur du maximum de vraisemblance aussi

```
1 > fitdistr(y,"poisson")
2     lambda
3     0.93229167
4     (0.04023135)
5 > logvrais = function(L){sum(log(dpois(y,L)))}
6 > optim(1,function(t) -logvrais(t))
7 $par
8 [1] 0.9322266
```

donc $\hat{\lambda} = 0.932$.

Test d'ajustement V

L'estimateur de la méthode des moments est $\hat{\lambda} = \bar{y}$

```
1 > (GF = goodfit(y,type="poisson"))
2
3 count observed      fitted pearson residual
4      0      229 226.7427226      0.14990574
5      1      211 211.3903507     -0.02684803
6      2       93  98.5387312     -0.55796481
7      3       35  30.6222793      0.79109619
8      4        7   7.1372240     -0.05136476
9      5        0   1.3307949     -1.15360083
10     6        0   0.2067815     -0.45473234
11     7        1   0.0275401      5.49264136
12 > summary(GF)
13
14      Goodness-of-fit test for poisson distribution
15
16                      X^2 df    P(> X^2)
17 Likelihood Ratio  9.262686  4  0.05485867
```

Test d'ajustement VI

L'estimateur de la méthode des moments est $\hat{\lambda} = \bar{y}$

On peut tenter 6 classes $\{0, 1, 2, 3, 4, 5+\}$, comme dans l'article

```
1 > observed <- c(n[1:5], sum(n[6:8]))
2 > names(observed) = c(0:4, "5+")
3 > observed
4   0    1    2    3    4   5+
5 229 211  93  35   7   1
6 > expected = c(dpois(0:4, lambda), 1-ppois(4, lambda))
7 > names(expected)=names(observed)
8 > expected
9 0     1     2     3     4     5+
10 0.39 0.37 0.17 0.05 0.01 0.00
11 > chisq.test(x=observed, p=expected)
12
13   Chi-squared test for given probabilities
14
15 data:  observed
16 X-squared = 1.1692, df = 5, p-value = 0.9478
```

Test d'ajustement VII

On peut tenter 5 classes $\{0, 1, 2, 3, 4+\}$,

```
1 > observed <- c(n[1:4], sum(n[5:8]))
2 > names(observed) = c(0:3, "4+")
3 > observed
4   0    1    2    3   4+
5 229 211  93  35    8
6 > expected = c(dpois(0:3, lambda), 1-ppois(3, lambda))
7 > names(expected) = c(0:3, "4+")
8 > expected
9   0    1    2    3   4+
10 0.39 0.37 0.17 0.05 0.02
11 > chisq.test(x=observed, p=expected)
12
13   Chi-squared test for given probabilities
14
15 data:  observed
16 X-squared = 1.0176, df = 4, p-value = 0.9071
```

Test d'ajustement VIII

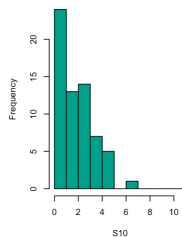
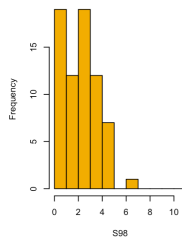
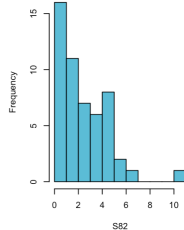
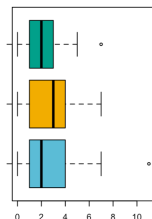
Exemple 2 : Sur trois coupes du monde de soccer, on analyse le nombre de buts par match. A-t-on des lois de Poisson ?

```
1 > soccer1982 = read.table("http://freakonometrics.free
    .fr/soccer1982")
2 > S82 = (soccer1982$V1+soccer1982$V2)
3 > soccer1998 = read.table("http://freakonometrics.free
    .fr/soccer1998")
4 > S98 = (soccer1998$V1+soccer1998$V2)
5 > soccer2010 = read.table("http://freakonometrics.free
    .fr/soccer2010")
6 > S10 = (soccer2010$V1+soccer2010$V3)
```

On notera $\mathbf{x} = \{x_1, \dots, x_{52}\}$, $\mathbf{y} = \{y_1, \dots, y_{70}\}$, $\mathbf{z} = \{z_1, \dots, z_{64}\}$ les trois échantillons

Test d'ajustement IX

```
1 > boxplot(S82,S98,S10,horizontal=TRUE)
2 > hist(S82,breaks=0:11)
```



```
1 > library(vcd)
2 > GF10 = goodfit(S10, type="poisson")
3 > summary(GF10)
```

Goodness-of-fit test for poisson distribution

	X ²	df	P(> X ²)
--	----------------	----	----------------------

Likelihood Ratio	5.586765	5	0.3485255
------------------	----------	---	-----------

Test d'ajustement X

En 2010, on ne rejette pas l'hypothèse (H_0) que le nombre de buts par match Z suit une loi de Poisson,

```
1 > GF10
2
3 Observed and fitted values for poisson distribution
4 with parameters estimated by 'ML'
5
6 count observed      fitted pearson residual
7      0         7  6.6409703          0.1393204
8      1        17 15.0459484          0.5037630
9      2        13 17.0442384         -0.9795981
10     3        14 12.8719508          0.3144169
11     4         7  7.2907534         -0.1076809
12     5         5  3.3036226          0.9333129
13     6         0  1.2474617         -1.1168982
14     7         1  0.4037543          0.5972266
```

Mais pour 1982, on la rejette

Test d'ajustement XI

```
1 > GF82=goodfit(S82,type="poisson")
2 > summary(GF82)
3
4     Goodness-of-fit test for poisson distribution
5
6               X^2 df    P(> X^2)
7 Likelihood Ratio 17.397  7 0.01500801
8 > GF82
9
10 Observed and fitted values for poisson distribution
11 with parameters estimated by 'ML'
12
13 count observed      fitted pearson residual
14    0         7  3.137892546      2.18024510
15    1         9  8.810236764      0.06393200
16    2        11 12.368216995     -0.38904643
17    3         7 11.575382572     -1.34480631
18    ...
19   11         1  0.006719346     10.61881076
```

Test d'ajustement ★★★

Si le test de Kolmogorov-Smirnov est l'outil adapté pour les variables continues, on peut aussi utiliser le test du chi-deux, en découpant en classes.

```
1 > x = Davis$height[Davis$sex=="F"]
2 > seuils=c(-Inf,160.5,162.5,165.5,169.5,Inf)
3 > x_cut = cut(x,seuils)
4 > table(x_cut)/length(x)
5 x_cut
6 (-Inf,160] (160,162] (162,166] (166,170] (170, Inf]
7          0.196      0.125      0.241      0.241      0.196
```

et on peut comparer avec ce que donnerait une loi F (ici une loi normale $\mathcal{N}(165, 5^2)$), car $p_j = F(s_{j-1}) - F(s_j)$ pour des seuils s_j .

```
1 > (prob = diff(pnorm(seuils,165,5)))
2 [1] 0.184 0.124 0.231 0.276 0.184
```

Test d'ajustement ★★★

On utilise alors le test du chi-deux

```
1 > chisq.test ( x = table(x_cut) , p = prob )
2
3   Chi-squared test for given probabilities
4
5 data:  table(x_cut)
6 X-squared = 0.7308, df = 4, p-value = 0.9475
```

(ce test est sensible aux seuils retenus)

Un peu de formalisme...

Tableau de contingence

X peut prendre les modalités $\{x_1, \dots, x_I\}$ et Y les modalités $\{y_1, \dots, y_J\}$. On appelle **tableau de contingence** la matrice N , $I \times J$, $N = [n_{ij}]$ où n_{ij} est le nombre d'individus dont les modalités sont x_i et y_j . On parle parfois aussi de **tri-croisé**.

Example Considérons l'exemple où X désigne la couleur des cheveux, et Y la couleur des yeux, de la base HairEyeColor,

```
1 > data(HairEyeColor)
2 > HairEyeColor[, , Sex="Female"]
3      Eye
4 Hair   Brown Blue Hazel Green
5   Black   36    9     5     2
6   Brown   66   34    29    14
7    Red   16    7     7     7
8   Blond    4   64    5     8
```

Un peu de formalisme...

Effets marginaux

Les **effets marginaux** sont notés

$$n_{i,\cdot} = \sum_j n_{i,j} \text{ et } n_{\cdot,j} = \sum_i n_{i,j}$$

L'effectif total de la population est alors

$$n = \sum_i n_{i,\cdot} = \sum_j n_{\cdot,j} = \sum_{i,j} n_{i,j}$$

```
1 > apply(HairEyeColor[, , Sex="Female"], 2, sum)
2 Brown   Blue  Hazel  Green
3    122    114     46     31
4 > apply(HairEyeColor[, , Sex="Female"], 1, sum)
5 Black Brown   Red  Blond
6     52    143     37     81
```

Un peu de formalisme...

On pose alors $F = \frac{1}{n}N = [f_{i,j}]$, où $f_{i,j} = \frac{n_{i,j}}{n}$.

```
1 > HairEyeColor[, , Sex="Female"] / sum(HairEyeColor[, , Sex  
    ="Female"])  
2      Eye  
3 Hair      Brown      Blue      Hazel      Green  
4   Black 0.11501597 0.02875399 0.01597444 0.006389776  
5   Brown 0.21086262 0.10862620 0.09265176 0.044728435  
6    Red  0.05111821 0.02236422 0.02236422 0.022364217  
7   Blond 0.01277955 0.20447284 0.01597444 0.025559105
```

De la même manière, on peut définir les effets marginaux

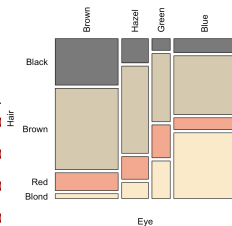
$$f_{i,\cdot} = \sum_j f_{i,j} \text{ et } f_{\cdot,j} = \sum_i f_{i,j}$$

Probabilités conditionnelles

	brown	hazel	green	blue	
black	63.0%	13.9%	4.6%	18.5%	100.0%
brown	41.6%	18.9%	10.1%	29.4%	100.0%
red	36.6%	19.7%	19.7%	23.9%	100.0%
blond	5.5%	7.9%	12.6%	74.0%	100.0%
	37.2%	15.7%	10.8%	36.3%	



	brown	hazel	green	blue	
black	30.9%	16.1%	7.8%	9.3%	18.2%
brown	54.1%	58.1%	45.3%	39.1%	48.3%
red	11.8%	15.1%	21.9%	7.9%	12.0%
blond	3.2%	10.8%	25.0%	43.7%	21.5%
	100.0%	100.0%	100.0%	100.0%	



Test d'indépendance I

Indépendance $X \perp\!\!\!\perp Y$

Soit X et Y deux variables discrètes, X et Y sont indépendantes - noté $X \perp\!\!\!\perp Y$ - si

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y], \quad \forall x, y.$$

Compte tenu des notations précédentes,

- ▶ on estime $\mathbb{P}[X = x_i, Y = y_j]$ par $\hat{p}_{i,j} = f_{i,j} = \frac{n_{i,j}}{n}$
- ▶ on estime $\mathbb{P}[X = x_i]$ par $\hat{p}_{i,\cdot} = f_{i,\cdot} = \frac{n_{i,\cdot}}{n}$
- ▶ on estime $\mathbb{P}[Y = y_j]$ par $\hat{p}_{\cdot,j} = f_{\cdot,j} = \frac{n_{\cdot,j}}{n}$

Test d'indépendance II

Indépendance empirique $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$

Soit $(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), \dots, (x_n, y_n)\}$ des couples de variables catégorielles appariées. Si $[n_{i,j}]$ est le tableau de contingence associé, on dira que \mathbf{x} et \mathbf{y} sont empiriquement indépendants - noté $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ - si

$$\hat{p}_{i,j} = \hat{p}_{i,\cdot} \hat{p}_{\cdot,j} \text{ ou } \frac{n_{i,j}}{n} = \frac{n_{i,\cdot}}{n} \frac{n_{\cdot,j}}{n}, \quad \forall i \text{ et } j$$

$$\text{ou } n_{i,j} = \frac{n_{i,\cdot} n_{\cdot,j}}{n}, \quad \forall i \text{ et } j$$

On pourra noter $\hat{p}_{i,j}^\perp = \hat{p}_{i,\cdot} \hat{p}_{\cdot,j}$, et on aura indépendance si $\mathbf{p} = \mathbf{p}^\perp$

Un test naturel sera un test du chi-deux.

Test d'indépendance III

Test $H_0 : X \perp\!\!\!\perp Y$ contre $H_1 : X \not\perp\!\!\!\perp Y$

Soit $(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), \dots, (x_n, y_n)\}$ des couples de variables catégorielles appariées. Si $[n_{i,j}]$ est le tableau de contingence associé, pour tester $H_0 : X \perp\!\!\!\perp Y$ contre $H_1 : X \not\perp\!\!\!\perp Y$ la statistique de test est

$$Q = \sum_{i=1}^I \sum_{j=1}^J \frac{(n\hat{p}_{i,j} - n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j})^2}{n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{i,j} - n_{i,\cdot}n_{\cdot,j}/n)^2}{n_{i,\cdot}n_{\cdot,j}/n}$$

Si $H_0 : X \perp\!\!\!\perp Y$ est vraie, $Q \sim \chi^2((I-1)(J-1))$. Et donc

► on rejette H_0 si $q > Q_{(I-1)(J-1)}^{-1}(1 - \alpha)$

où Q_ν est la fonction de répartition de la loi du chi-deux, $\chi^2(\nu)$.

Test d'indépendance IV

Notons qu'on peut aussi écrire la statistique de test sur les probabilités, et pas les comptages

$$Q = \sum_{i=1}^I \sum_{j=1}^J \frac{(n\hat{p}_{i,j} - n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j})^2}{n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}} = n \sum_{i=1}^I \sum_{j=1}^J \frac{(\hat{p}_{i,j} - \hat{p}_{i,\cdot}\hat{p}_{\cdot,j})^2}{\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}}$$

On peut également écrire

$$Q = \sum_{i=1}^I \sum_{j=1}^J \epsilon_{i,j}^2 \text{ où } \epsilon_{i,j} = \frac{(n\hat{p}_{i,j} - n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j})}{\sqrt{n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}}}$$

où, si $H_0 : X \perp\!\!\!\perp Y$ est vraie, $\epsilon_{i,j} \approx \mathcal{N}(0, 1)$.

$\epsilon_{i,j}^2$ est appelée **contribution au test du chi-deux**

Test d'indépendance V

```
1 > N = HairEyeColor[, , Sex="Female"] + HairEyeColor[, , Sex
  = "Male"]
2 > (Q = chisq.test(N))
3
4 Pearson's Chi-squared test
5
6 data:  N
7 X-squared = 138.29, df = 9, p-value < 2.2e-16
8 > Q$observed
9      Eye
10 Hair   Brown Blue Hazel Green
11 Black    68   20    15     5
12 Brown   119   84    54    29
13 Red     26   17    14    14
14 Blond     7   94    10    16
```

Test d'indépendance VI

```
1 > Q$expected
2      Eye
3 Hair      Brown      Blue      Hazel      Green
4  Black  40.13514  39.22297  16.96622  11.675676
5  Brown 106.28378 103.86824  44.92905  30.918919
6  Red    26.38514  25.78547  11.15372   7.675676
7  Blond  47.19595  46.12331  19.95101  13.729730
```

Comme attendu, on notera que $n_{\cdot j}^{\perp} = n_{\cdot j}$ pour tout j

```
1 > apply(Q$observed, 2, sum)
2 Brown  Blue  Hazel  Green
3   220   215    93    64
4 > apply(Q$expected, 2, sum)
5 Brown  Blue  Hazel  Green
6   220   215    93    64
```

(et on vérifiera que $n_{\cdot j}^{\perp} = n_{\cdot j}$ pour tout j)

Test d'indépendance VII

```
1 > Q
2
3   Pearson's Chi-squared test
4
5 data:  N
6 X-squared = 138.29, df = 9, p-value < 2.2e-16
```

On rejette ici $H_0 : X \perp\!\!\!\perp Y$ car q dépasse le quantile à 95% d'une loi du $\chi^2(3 \times 3)$,

```
1 > qchisq(.95, 3*3)
2 [1] 16.91898
```

avec une p -value inférieure à 10^{-16} .

On peut aussi calculer les **résidus** $\epsilon_{i,j} = \frac{(n\hat{p}_{i,j} - n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j})}{\sqrt{n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j}}}$

Test d'indépendance VIII

Les résidus sont $\epsilon_{i,j} = \frac{(n\hat{p}_{i,j} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})}{\sqrt{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}}$

```
1 > Q$residuals
2      Eye
3 Hair      Brown      Blue      Hazel      Green
4  Black  4.3984 -3.0694 -0.4774 -1.9537
5  Brown  1.2335 -1.9495  1.3533 -0.3451
6  Red    -0.0750 -1.7301  0.8523  2.2827
7  Blond -5.8510  7.0496 -2.2278  0.6127
```

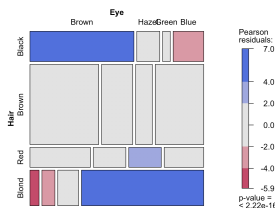
On rejette H_0 car on a

- ▶ trop de personnes aux cheveux Black ayant les yeux Brown
- ▶ trop de personnes aux cheveux Blond ayant les yeux Blue
- ▶ pas assez de personnes Black ayant les yeux Blue
- ▶ pas assez de personnes Blond ayant les yeux Brown

Test d'indépendance IX

	brown	hazel	green	blue	
black	68	15	5	20	108
brown	119	54	29	84	286
red	26	14	14	17	71
blond	7	10	16	94	127
	220	93	64	215	

	brown	hazel	green	blue	
black	40	17	12	39	108
brown	106	45	31	104	286
red	26	11	8	26	71
blond	47	20	14	46	127
	220	93	64	215	



on compare $n_{i,j}$ et $n_{i,j}^\perp$

$$n_{i,j}^\perp = \frac{n_{i,\cdot} \cdot n_{\cdot,j}}{n}$$

Test d'indépendance I

Exemple 3 : En analysant les données relatives à la peine de mort pour les condamnées pour meurtre en Floride 1976-1987, on a les statistiques suivantes

- ▶ meurtrier de “race blanche” et victime de “race blanche”: 53 condamnés à mort, 414 non condamnés à mort
- ▶ meurtrier de “race blanche” et victime de “race noire”: 0 condamné à mort, 16 non condamnés à mort
- ▶ meurtrier de “race noire” et victime de “race blanche”: 11 condamnés à mort, 37 non condamnés à mort
- ▶ meurtrier de “race noire” et victime de “race noire”: 4 condamnés à mort, 139 non condamnés à mort

Que peut-on dire (statistiquement) sur la base de ces statistiques ?

Test d'indépendance II

Indépendance entre la "race" de la victime et la condamnation

```
1 > N = matrix(c(53+11,0+4,414+37,139+16),2,2)
2 > rownames(N) = c("blanc","noir")
3 > colnames(N) = c("a mort","pas a mort")
4 > Q = chisq.test(N)
5 > Q$observed
6      a mort pas a mort
7 blanc      64      451
8 noir       4      155
9 > Q$expected
10     a mort pas a mort
11 blanc 51.95846 463.0415
12 noir  16.04154 142.9585
13 > Q$residuals
14     a mort pas a mort
15 blanc  1.670529 -0.5595929
16 noir   -3.006485  1.0071107
17 > Q
18 X-squared = 12.087, df = 1, p-value = 0.0005077
```

Test d'indépendance III

Indépendance entre la "race" de l'accusé(e) et la condamnation

```
1 > N = matrix(c(53+0,11+4,414+16,139+37),2,2)
2 > rownames(N) = c("blanc","noir")
3 > colnames(N) = c("a mort","pas a mort")
4 > Q = chisq.test(N)
5 > Q$observed
6      a mort pas a mort
7 blanc      53      430
8 noir      15      176
9 > Q$expected
10     a mort pas a mort
11 blanc 48.72997      434.27
12 noir  19.27003      171.73
13 > Q$residuals
14     a mort pas a mort
15 blanc  0.6116920 -0.2049042
16 noir  -0.9727242  0.3258426
17 > Q
18 X-squared = 1.1447, df = 1, p-value = 0.2847
```

Test d'indépendance IV

Indépendance entre la "race" de la victime et l'accusé(e)

```
1 > N = matrix(c(53+114,0+16,11+37,139+4),2,2)
2 > rownames(N) = c("blanc","noir")
3 > colnames(N) = c("blanc","blanc")
4 > Q = chisq.test(N)
5 > Q$observed
6      blanc blanc
7 blanc   167    48
8 noir    16   143
9 > Q$expected
10      blanc      blanc
11 blanc 105.20053 109.79947
12 noir   77.79947  81.20053
13 > Q$residuals
14      blanc      blanc
15 blanc  6.025259 -5.897726
16 noir  -7.006424  6.858123
17 > Q
18 X-squared = 164.52, df = 1, p-value < 2.2e-16
```