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 $\{x_1, \cdots, x_J\}$. On appelle tableau de comptage le vecteur $\boldsymbol{n}$ de taille J $\boldsymbol{n} = [n_j] = (n_1, \cdots, n_J)$ 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`,

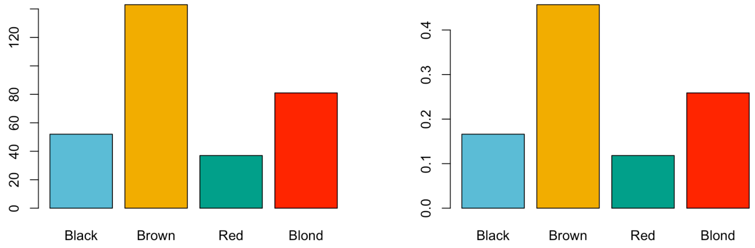
```
> data(HairEyeColor)
> n = apply(HairEyeColor[,,Sex="Female"],1,sum)
> n
Black Brown   Red Blond
   52   143    37    81
> n/sum(n)
    Black     Brown       Red     Blond
0.1661342 0.4568690 0.1182109 0.2587859
```

Un peu de formalisme...

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

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

```
> barplot(n)
> f = n/sum(n)
> barplot(f)
```



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

Loi multinomiale

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

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

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

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

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

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

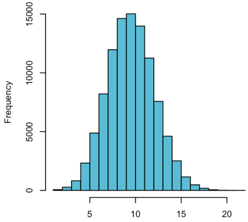
(admis)

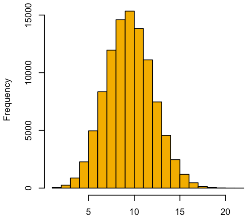
Loi multinomiale

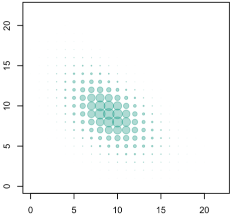
Les variables N_i et N_j ne sont pas indépendantes, car $\displaystyle\sum_{j=1}^{J} N_j = n$

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

```
> X = sample(1:3,size=n,replace=TRUE)
> 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 $\boldsymbol{n} = (n_1, \cdots, n_J)$ tel que $n_1 + \cdots + n_J = n$.

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 $\{x_1, \cdots, x_n\}$ est une collection de variables catégorielles indépendantes, de probabilités $\boldsymbol{p} = (p_1, \cdots, p_J)$, 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)

Test

Cette dernière propriété permet de proposer un test de fréquence

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

Si $\{x_1, \cdots, x_n\}$ est une collection de variables catégorielles indépendantes, de probabilités $\boldsymbol{p} = (p_1, \cdots, p_J)$, pour tester $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ contre $H_1 : \boldsymbol{p} \neq \boldsymbol{p}_0$ 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 $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ 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 ?

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

$$q = \frac{(88-100)^2}{100} + \frac{(109-100)^2}{100} + \frac{(107-100)^2}{100} + \frac{(94-100)^2}{100} + \frac{(105-}{1}$$

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

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

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

et la p-value vaut 36.7%

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

Loi multinomiale, test

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

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

$$q = \frac{(89-100)^2}{100} + \frac{(131-100)^2}{100} + \frac{(93-100)^2}{100} + \frac{(92-100)^2}{100} + \frac{(104-}{10}$$

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

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

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

et la p-value vaut 36.7%

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

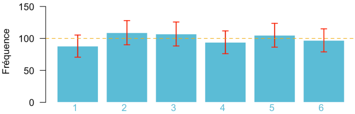
Tests multiples ★★★

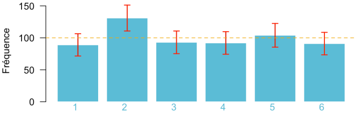
On a ponctuellement des intervalles de confidance, sur les probabilités

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

et sur les fréquences

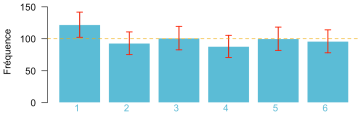
$$\left[n_j \pm u_{1-\alpha} \cdot \sqrt{\frac{n_j(n-n_j)}{n}}\right]$$

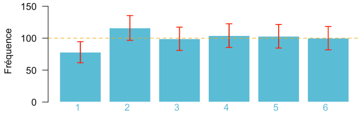




Tests multiples ★★★

Ces intervalles de confiance sont associés à 6 tests simples



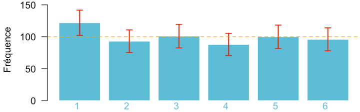


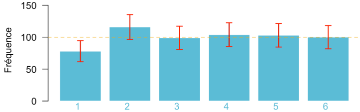
On peut refuser un test simple (un sur les six)

```
> table(X)
  1   2   3   4   5   6
 78 116  99 104 103 100
> prop.test(table(X)[1],600,1/6)

  1-sample proportions test with continuity correction

data:  table(X)[1] out of 600, null probability 1/6
X-squared = 5.547, df = 1, p-value = 0.01851
alternative hypothesis: true p is not equal to
    0.1666667
95 percent confidence interval:
 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, $H_0 : p_1 = \cdots = p_6$.

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

i.e.

```
> chisq.test(table(X), p = rep(1/6,6))

  Chi-squared test for given probabilities

data:  table(X)
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

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

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

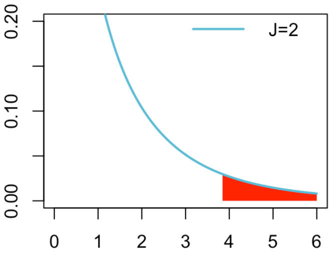
$$q = \sum_{j=1}^{J} \frac{(N\hat{p}_{0,j} - Np_{0,j})^2}{Np_{0,j}} = \sum_{j=1}^{J} \frac{(n_j - Np_{0,j})^2}{Np_{0,j}}$$

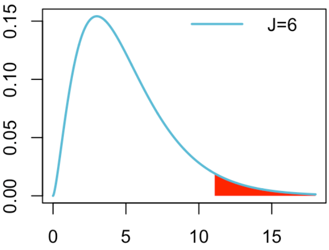
la p-value est

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

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 $\mathcal{P}(2)$

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

On pourra noter

```
> p = c(dpois(0:4,2),1-ppois(4,2))
> names(p) = c(0:4,"5+")
> p
    0     1     2     3     4    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 $\{x_1, \cdots, x_n\}$ est une collection de variables indépendantes de loi f_0. Notons $\boldsymbol{p}_0 = (p_{0,1}, \cdots, p_{0,J})$ 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 $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ 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

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

Test d'ajustement IV

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

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

L'estimateur du maximum de vraisemblance aussi

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

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

Test d'ajustement V

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

```
> (GF = goodfit(y,type="poisson"))

 count  observed       fitted  pearson residual
     0       229  226.7427226        0.14990574
     1       211  211.3903507       -0.02684803
     2        93   98.5387312       -0.55796481
     3        35   30.6222793        0.79109619
     4         7    7.1372240       -0.05136476
     5         0    1.3307949       -1.15360083
     6         0    0.2067815       -0.45473234
     7         1    0.0275401        5.49264136
> summary(GF)

   Goodness-of-fit test for poisson distribution

                       X^2 df    P(> X^2)
Likelihood Ratio 9.262686  4 0.05485867
```

Test d'ajustement VI

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

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

```
> observed <- c(n[1:5],sum(n[6:8]))
> names(observed) = c(0:4,"5+")
> observed
  0   1   2   3   4  5+
229 211  93  35   7   1
> expected = c(dpois(0:4,lambda),1-ppois(4,lambda))
> names(expected)=names(observed)
> expected
0     1     2     3     4    5+
0.39 0.37 0.17 0.05 0.01 0.00
> chisq.test(x=observed, p=expected)

  Chi-squared test for given probabilities

data:  observed
X-squared = 1.1692, df = 5, p-value = 0.9478
```

Test d'ajustement VII

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

```
> observed <- c(n[1:4],sum(n[5:8]))
> names(observed) = c(0:3,"4+")
> observed
  0   1   2   3  4+
229 211  93  35   8
> expected = c(dpois(0:3,lambda),1-ppois(3,lambda))
> names(expected) = c(0:3,"4+")
> expected
   0    1    2    3   4+
0.39 0.37 0.17 0.05 0.02
> chisq.test(x=observed, p=expected)

  Chi-squared test for given probabilities

data:  observed
X-squared = 1.0176, df = 4, p-value = 0.9071
```

Test d'ajustement VIII

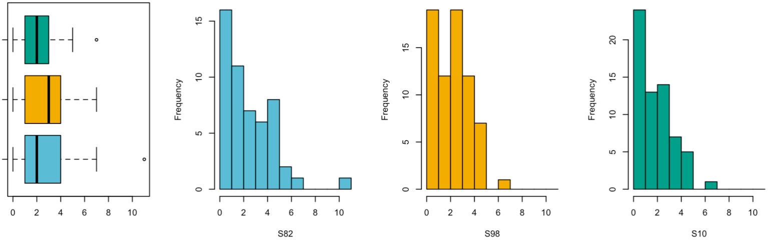
Example 2 : Sur trois coupes du monde de soccer, on analyse le nombre de buts par match. A-t-on des lois de Poisson ?

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

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

Test d'ajustement IX

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




```
> library(vcd)
> GF10 = goodfit(S10, type="poisson")
> 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,

```
> GF10

Observed and fitted values for poisson distribution
with parameters estimated by ‘ML’

 count observed     fitted pearson residual
     0        7  6.6409703        0.1393204
     1       17 15.0459484        0.5037630
     2       13 17.0442384       -0.9795981
     3       14 12.8719508        0.3144169
     4        7  7.2907534       -0.1076809
     5        5  3.3036226        0.9333129
     6        0  1.2474617       -1.1168982
     7        1  0.4037543        0.5972266
```

Mais pour 1982, on la rejette

Test d'ajustement XI

```
> GF82=goodfit(S82,type="poisson")
> summary(GF82)

   Goodness-of-fit test for poisson distribution

                     X^2 df    P(> X^2)
Likelihood Ratio 17.397  7 0.01500801
> GF82

Observed and fitted values for poisson distribution
with parameters estimated by ‘ML’

 count observed       fitted pearson residual
     0        7  3.137892546       2.18024510
     1        9  8.810236764       0.06393200
     2       11 12.368216995      -0.38904643
     3        7 11.575382572      -1.34480631
    ...
    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.

```
> x = Davis$height[Davis$sex=="F"]
> seuils=c(-Inf,160.5,162.5,165.5,169.5,Inf)
> x_cut = cut(x,seuils)
> table(x_cut)/length(x)
x_cut
(-Inf,160]  (160,162]  (162,166]  (166,170] (170, Inf]
     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.

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

Test d'ajustement ★★★

On utilise alors le test du chi-deux

```
> chisq.test ( x = table(x_cut) , p = prob )

  Chi-squared test for given probabilities

data:  table(x_cut)
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 $\{x_1, \cdots, x_I\}$ et Y les modalités $\{y_1, \cdots, y_J\}$. On appelle tableau de contingence la matrice N, $I \times J$, $N = [n_{i,j}]$ où $n_{i,j}$ 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,

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

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

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

Un peu de formalisme...

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

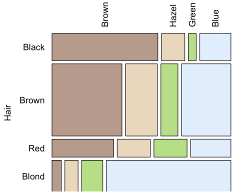
```
> HairEyeColor[,,Sex="Female"]/sum(HairEyeColor[,,Sex
    ="Female"])
       Eye
Hair          Brown       Blue      Hazel       Green
  Black 0.11501597 0.02875399 0.01597444 0.006389776
  Brown 0.21086262 0.10862620 0.09265176 0.044728435
  Red   0.05111821 0.02236422 0.02236422 0.022364217
  Blond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

De la même manière, on peut définir les effets marginaux

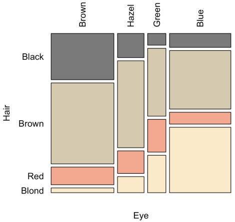
$$f_{i,\cdot} = \sum_j f_{i,j} \text{ et } f_{\cdot,j} = \sum_i f_{i,j}$$

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

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

Compte tenu des notations précédentes,

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

Test d'indépendance II

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

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

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

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

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

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 $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ des couples de variables catégorielles appareillées. 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

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

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

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

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

```
> N = HairEyeColor[,,Sex="Female"]+HairEyeColor[,,Sex
    ="Male"]
> (Q = chisq.test(N))

  Pearson's Chi-squared test

data:  N
X-squared = 138.29, df = 9, p-value < 2.2e-16
> Q$observed
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
```

Test d'indépendance VI

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

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

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

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

Test d'indépendance VII

```
> Q

  Pearson's Chi-squared test

data:  N
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 $\chi^2(3 \times 3)$,

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

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

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

Test d'indépendance VIII

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

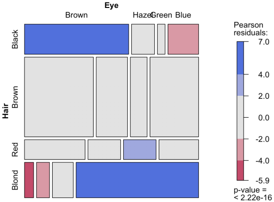
```
> Q$residuals
       Eye
Hair       Brown    Blue   Hazel   Green
  Black   4.3984 -3.0694 -0.4774 -1.9537
  Brown   1.2335 -1.9495  1.3533 -0.3451
  Red    -0.0750 -1.7301  0.8523  2.2827
  Blond  -5.8510  7.0496 -2.2278  0.6127
```

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

Example 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

```
> N = matrix(c(53+11,0+4,414+37,139+16),2,2)
> rownames(N) = c("blanc","noir")
> colnames(N) = c("a mort","pas a mort")
> Q = chisq.test(N)
> Q$observed
      a mort pas a mort
blanc     64        451
noir       4        155
> Q$expected
        a mort pas a mort
blanc 51.95846   463.0415
noir  16.04154   142.9585
> Q$residuals
         a mort pas a mort
blanc  1.670529 -0.5595929
noir  -3.006485  1.0071107
> Q
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

```
> N = matrix(c(53+0,11+4,414+16,139+37),2,2)
> rownames(N) = c("blanc","noir")
> colnames(N) = c("a mort","pas a mort")
> Q = chisq.test(N)
> Q$observed
      a mort pas a mort
blanc     53        430
noir      15        176
> Q$expected
        a mort pas a mort
blanc 48.72997     434.27
noir  19.27003     171.73
> Q$residuals
          a mort pas a mort
blanc  0.6116920 -0.2049042
noir  -0.9727242  0.3258426
> Q
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)

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

Exemple de tests

EXEMPLE 2

Un chercheur désire comparer la distribution des revenus des familles immigrantes du Québec à celle des revenus de l'ensemble des familles québécoises. Cette dernière distribution est connue grâce au recensement, mais pas celle des revenus des familles immigrantes du Québec. Le chercheur décide donc de procéder par échantillonnage pour faire son étude. En prenant un échantillon aléatoire de 500 familles immigrantes, il obtient la distribution suivante.

Répartition d'un échantillon de 500 familles immigrantes selon la tranche de revenu

Revenu (milliers $)	Moins de 25	[25 ; 50[	[50 ; 75[	[75 ; 100[	100 et plus	Total
Nombre de familles	44	142	129	65	120	**500**
Pourcentage	8,8 %	28,4 %	25,8 %	13,0 %	24,0 %	**100 %**

Répartition des familles selon la tranche de revenu, Québec, 2011

Revenu (milliers $)	Moins de 25	[25 ; 50[	[50 ; 75[	[75 ; 100[	100 et plus	Total
Pourcentage	6,1 %	26,3 %	23,4 %	16,4 %	27,8 %	**100 %**

Source : Statistique Canada. *Tableau 202-0408*, CANSIM, juin 2013.

(via Simard (2015))

Exemple de tests

> En comparant les pourcentages des deux distributions, on constate que les familles immigrantes sont moins riches : un plus grand pourcentage de ces familles ont un revenu faible et un plus petit pourcentage ont un revenu élevé.
>
> En fait, cette affirmation est vraie pour les 500 familles immigrantes de notre échantillon, mais est-elle vraie pour l'ensemble de toutes les familles immigrantes du Québec ? Il est en effet possible que les distributions pour l'ensemble de toutes les familles immigrantes et québécoises soient identiques et que les écarts observés ci-dessus soient attribuables à la variation d'échantillonnage causée par le hasard. Un test d'ajustement du khi-deux permet de savoir si c'est le cas.
>
> Effectuer un test d'ajustement du khi-deux, au seuil de signification de 0,01, pour déterminer si la distribution des revenus des familles immigrantes est identique à celle des revenus des familles québécoises.

(via Simard (2015))

```
> x = c(44, 142, 129, 65, 120)
> p = c(6.1, 26.3, 23.4, 16.4, 27.8)/100
> sum( (x-500*p)^2/(500*p) )
[1] 14.16609
```

Exemple de tests

```
> chisq.test(x=x, p=p)

  Chi-squared test for given probabilities

data:  c(44, 142, 129, 65, 120)
X-squared = 14.166, df = 4, p-value = 0.006783
```

La p-value vaut 0.006783 < 5% donc on rejette H_0

Exemple de tests

EXEMPLE

On désire mesurer l'effet des nouvelles technologies de communication sur la vie quotidienne des Québécois de 25-64 ans. Pour ce faire, on prélève au hasard un échantillon de 800 personnes dans cette population. Comme on considère que le niveau de scolarité est une variable importante dans ce genre d'étude, on veut s'assurer de la représentativité de l'échantillon pour cette variable avant de procéder à la cueillette des données. Les statistiques présentées dans les deux tableaux suivants permettent-elles d'affirmer que l'échantillon est représentatif des Québécois de 25-64 ans en ce qui concerne le niveau de scolarité, au seuil de signification de 0,05 ?

Répartition des Québécois de 25-64 ans selon le plus haut niveau de scolarité atteint, Québec, 2012

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	Total
Pourcentage	12,3 %	33,5 %	22,2 %	32,0 %	100,0 %

Source : Statistique Canada. *Enquête sur la population active,* 2013, adapté par l'Institut de la statistique du Québec, juin 2014.

Répartition des 800 répondants selon le niveau de scolarité

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	Total
Effectifs	91	258	207	244	800

(via Simard (2015))

Exemple de tests

Répartition des Québécois de 25-64 ans selon le plus haut niveau de scolarité atteint, Québec, 2012

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	**Total**
Pourcentage	12,3 %	33,5 %	22,2 %	32,0 %	**100,0 %**

Source : Statistique Canada. *Enquête sur la population active,* 2013, adapté par l'Institut de la statistique du Québec, juin 2014.

Répartition des 800 répondants selon le niveau de scolarité

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	**Total**
Effectifs	91	258	207	244	**800**

(via Simard (2015))

```
> x = c(91, 258, 207, 244)
> p = c(12.3, 33.5, 22.2, 32)/100
> sum( (x-800*p)^2/(800*p) )
[1] 6.35903
```

Exemple de tests

```
> chisq.test(x = x, p = p)

  Chi-squared test for given probabilities

data:  x
X-squared = 6.359, df = 3, p-value = 0.09539
```

Exemple de tests

EXEMPLE

Pour dresser le profil statistique des passagers des navires de croisières qui accostent au port de Québec, on prélève un échantillon aléatoire de 1 000 croisiéristes. La moyenne d'âge de ces derniers est de 64,7 ans avec un écart type corrigé de 12,1 ans. Le tableau suivant donne la distribution de l'âge des personnes de l'échantillon. Au seuil de signification de 0,05, ces données permettent-elles d'affirmer que la distribution de l'âge des croisiéristes dans la population suit une distribution normale ?

Répartition des croisiéristes de l'échantillon selon l'âge

Âge (en ans)	[30; 40[	[40; 50[	[50; 60[	[60; 70[	[70; 80[	[80; 90[	Total
Effectifs	24	90	230	310	239	107	**1 000**

(via Simard (2015))

```
> x = c(24, 90, 230, 310, 239, 107)
> (p = diff(pnorm(seuils,64.7,12.1)))
[1] 0.02061 0.09160 0.23664 0.32046 0.22766 0.10303
> chisq.test(x = x, p = p)

  Chi-squared test for given probabilities

data:  x
X-squared = 1.8319, df = 5, p-value = 0.8719
```