Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

Rappels #4.3 (tests & loi multinomiale)

Lois Binomiales & Multinomiales

```
> library(car)
> model.matrix( ~ type - 1, data=Prestige ))
                       type  typebc  typeprof  typewc
gov.administrators     prof       0         1       0
general.managers       prof       0         1       0
nursing.aides            bc       1         0       0
physio.therapsts       prof       0         1       0
pharmacists            prof       0         1       0
medical.technicians      wc       0         0       1
radio.tv.announcers      wc       0         0       1
```

Convert $X \in \{C_1, \cdots, C_d\}$ into a d dimensional vector $\{0,1\}^d$,
$\mathbf{Y} = (\mathbf{1}_{C_1}, \mathbf{1}_{C_2}, \cdots, \mathbf{1}_{C_d})$, with $\mathbf{Y} =\sim \mathcal{M}(\mathbf{p})$ where $\mathbf{p} = (p_1, \cdots, p_d)$

We want to test $H_0 : \mathbf{p} = \mathbf{p}_0$ from a sample $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$

Lois Binomiales & Multinomiales

$\mathbf{Y} = (Y_1, \cdots, Y_d) \sim \mathcal{M}(\mathbf{p})$ où $\mathbf{p} = (p_1, \cdots, p_d)$ si

$$Y_1 + \cdots + Y_d = 1 \text{ et } Y_j \sim \mathcal{B}(p_j),\ \forall j \in \{1, \cdots, d\}$$

i.e. $\mathbf{Y} = (\mathbf{1}_{C_1}, \mathbf{1}_{C_2}, \cdots, \mathbf{1}_{C_d})$

$\mathbf{Y} = (Y_1, \cdots, Y_d) \sim \mathcal{M}(n, \mathbf{p})$ où $\mathbf{p} = (p_1, \cdots, p_d)$ si

$$Y_1 + \cdots + Y_d = n \text{ et } Y_j \sim \mathcal{B}(n, p_j),\ \forall j \in \{1, \cdots, d\}$$

cf loi multinomiale. Pour
$(y_1, \cdots, y_d) \in \mathcal{S}_{d,n} = \{(y_1, \cdots, y_d) \in \mathbb{N}^d : (y_1 + \cdots + y_d = n\}$

$$\mathbb{P}[(Y_1, \cdots, Y_d) = (y_1, \cdots, y_d)] = \frac{n!}{y_1! \cdots y_d!} p_1^{y_1} \cdots p_d^{y_d}$$

Example: $\mathbf{Y} = (Y_0, Y_1) \sim \mathcal{M}(n, \mathbf{p})$ où $\mathbf{p} = (p_0, p_1)$.

Lois Multinomiales : Inférence

$\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ i.i.d de loi $\mathcal{M}(\mathbf{p})$, alors

$$\mathcal{L}(\mathbf{p}; \mathbf{y}) = \prod_{i=1}^{n} \mathbb{P}(\mathbf{Y}_i = \mathbf{y}_i) = \prod_{i=1}^{n} \prod_{j=1}^{d} p_j^{y_{i,j}}$$

sous la contrainte que $\mathbf{p}^\top \mathbf{1} = 1$.
Posons $\mathbf{x} = (\mathbf{x}_{(d)}, x_d)$, i.e. $p_d = 1 - \mathbf{p}_{(d)}^\top \mathbf{1}$

$$\mathcal{L} = \prod_{i=1}^{n} \mathbb{P}(\mathbf{Y}_i = \mathbf{y}_i) = \prod_{i=1}^{n} \left(\prod_{j=1}^{d-1} p_j^{y_{i,j}} \right) \left(1 - \mathbf{p}_{(d)}^\top \mathbf{1}\right)^{\mathbf{y}_{i(d)}^\top \mathbf{1}}$$

La $j$ème condition du premier ordre est, si $s_j = \sum_{i=1}^{n} y_{i,j}$

$$\left.\frac{\partial \log \mathcal{L}}{\partial p_j}\right|_{\mathbf{p}=\widehat{\mathbf{p}}} = \frac{s_j}{\widehat{p}_j} - \frac{n - \mathbf{s}_{(d)}^\top \mathbf{1}}{1 - \widehat{\mathbf{p}}_{(d)}^\top \mathbf{1}} = 0, \text{ i.e. } \widehat{p}_j = \frac{s_j}{n}.$$

Lois Multinomiales : Inférence

L'estimateur du maximum de vraisemblance est

$$\widehat{\mathbf{p}} = (\widehat{p}_1, \cdots, \widehat{p}_d) = \left(\frac{s_1}{n}, \cdots, \frac{s_d}{n}\right)$$

Propriété: $\mathbb{E}(\widehat{\mathbf{p}}) = \mathbf{p}$ et $\text{Var}(\widehat{\mathbf{p}}) = \dfrac{1}{n}\boldsymbol{\Omega}$, où

$$\boldsymbol{\Omega} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_d \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_d \\ \vdots & \vdots & \ddots & \vdots \\ -p_dp_1 & -p_dp_2 & \cdots & p_d(1-p_d) \end{pmatrix}$$

$$\sqrt{n}\big(\widehat{\mathbf{p}} - \mathbf{p}\big) \overset{\mathcal{L}}{\to} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}),$$

Remarque $\text{rang}(\boldsymbol{\Omega}) = d - 1$.

Lois Multinomiales : Test

Test de Pearson (ou χ^2): $H_0 : \mathbf{p} = \mathbf{p}_0$, on utilise

$$Q = \sum_{j=1}^{d} \frac{(S_j - np_{0,j})^2}{np_{0,j}} \xrightarrow{\mathcal{L}} \chi^2(d-1),\ n \to \infty,$$

si H_0 est vraie, cf test du chi-deux.

On retrouvera ce test comme test d'indépendance.

χ^2-test

$$X^2 = \sum_{j=1}^{k} \frac{\big(\text{observed number of } i) - (\text{expected number of } i)\big)^2}{(\text{expected number of } i)}$$

compare with χ^2_{k-1}

	dice value					
	1	2	3	4	5	6
observed	4	6	17	16	8	9
expected	10	10	10	10	10	10

Hypothesis (H_0): dice is fair (against $(H1)$ dice is unfair)

$$X^2 = \frac{6^2}{10} + \frac{4^2}{10} + \frac{7^2}{10} + \frac{6^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} \simeq 14.2$$

The probability of getting a probability of 14.2 with a χ^2_5 is 1.4%

$$X^2 = \sum_{j=1}^{k} \frac{\big(\text{observed number of } i) - (\text{expected number of } i)\big)^2}{(\text{expected number of } i)}$$

	observed		total	expected ($\perp$)	
	men	women		men	women
right-handed	934	1070	2004	956	1048
left-handed	113	92	205	98	107
ambidextrous	20	8	28	13	15
total	1067	1170	2237	1067	1170

$$n \cdot \mathbb{P}(N^{\perp}_{rm}) = n \cdot \mathbb{P}(N_r)\mathbb{P}(N_m) = n\frac{n_r}{n}\frac{n_m}{n} = 2237\frac{2004}{2237}\frac{1067}{2237} \simeq 956$$

Hypothesis: left-handedness equally common for men and women

$$X^2 = \frac{22^2}{956} + \frac{22^2}{1048} + \frac{15^2}{98} + \frac{15^2}{107} + \frac{7^2}{13} + \frac{7^2}{15} \simeq 12$$

The probability of getting a probability of 12 with a χ^2_2 is 0.2%