

Modèles Linéaires Appliqués / Régression

Variable y catégorielle

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 1



Lois Binomiales & Multinomiales

$Y \sim \mathcal{B}(p)$:

$$\mathbb{P}[Y = y] = p^y(1-p)^{1-y} \begin{cases} p & \text{si } y = 1 \\ 1-p & \text{si } y = 0 \end{cases}, \text{ où } y \in \{0, 1\}$$

cf **loi de Bernoulli**, où $p = \mathbb{P}[Y = 1] = \mathbb{E}[Y] \in [0, 1]$.

$Y \sim \mathcal{B}(n, p)$:

$$\mathbb{P}[Y = y] = \binom{n}{y} p^y(1-p)^{n-y} \text{ où } y \in \{0, 1, 2, \dots, n\}$$

cf **loi binomiale**, où $\mathbb{E}[Y] = np$.

Y_1, \dots, Y_n i.i.d. $\mathcal{B}(p)$ alors $Y = \sum_{i=1}^n Y_i \sim \mathcal{B}(n, p)$

Lois Binomiales & Multinomiales

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

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

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

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

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

cf loi multinomiale. Pour

$$(y_1, \dots, y_d) \in \mathcal{S}_{d,n} = \{(y_1, \dots, y_d) \in \mathbb{N}^d : (y_1 + \dots + y_d = n)\}$$

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

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

Lois Binomiales : Inférence

y_1, y_2, \dots, y_n i.i.d de loi $\mathcal{B}(p)$, alors

$$\mathcal{L}(p; \mathbf{y}) = \prod_{i=1}^n \mathbb{P}(Y_i = y_i) = \prod_{i=1}^n p^{y_i} [1 - p]^{1-y_i}$$

et la **log-vraisemblance** est

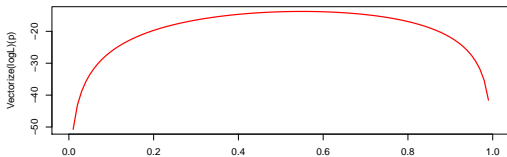
$$\log \mathcal{L}(p; \mathbf{y}) = \sum_{i=1}^n y_i \log[p] + (1 - y_i) \log[1 - p]$$

La condition du premier ordre est

$$\left. \frac{\partial \log \mathcal{L}(p; \mathbf{y})}{\partial p} \right|_{p=\hat{p}} = \sum_{i=1}^n \left(\frac{y_i}{\hat{p}} - \frac{1 - y_i}{1 - \hat{p}} \right) = 0, \text{ i.e. } \hat{p} = \frac{1}{n} \sum_{i=1}^n y_i$$

Lois Binomiales : Inférence

```
1 > set.seed(1)
2 > n=20
3 > (Y=sample(0:1,size=n,replace=TRUE))
4 [1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1
5 > (pn = mean(Y))
6 [1] 0.55
7 > p=seq(0,1,by=.01)
8 > neglogL = function(p){-sum(log(dbinom(Y,1,p)))}
9 > plot(p,-Vectorize(neglogL)(p))
10 > pml = optim(fn=neglogL,par=.5,method="BFGS")$par
11 [1] 0.5499996
```



Lois Binomiales : Inférence

Propriété du maximum de vraisemblance

$$\sqrt{n}(p - \hat{p}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(p))$$

où $I(p)$ est l'information de Fisher, i.e.

$$I(p) = -\mathbb{E} \left[\frac{\partial^2}{\partial p^2} \log f(Y, p) \right] = \frac{1}{p[1-p]}$$

$$\sqrt{n} \frac{p - \hat{p}}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ et } \sqrt{n} \frac{p - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

d'où un intervalle de confiance approché (à 95%) pour p de la forme

$$\left[\hat{p} \pm \frac{1.96}{\sqrt{n}} \sqrt{\hat{p}[1-\hat{p}]} \right].$$

Lois Binomiales : Inférence

On peut aussi construire un intervalle de confiance, par le théorème central limite, car $\hat{p} = \bar{y}$. On sait que

$$\sqrt{n} \frac{\bar{Y} - \mathbb{E}(Y)}{\sqrt{\text{Var}(Y)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1),$$

avec ici $\bar{Y} = \hat{p}$, $\mathbb{E}(Y) = p$ et $\text{Var}(Y) = p(1 - p)$, i.e. un intervalle de confiance est obtenu par l'approximation

$$\sqrt{n} \frac{\hat{p} - p}{\sqrt{\hat{p}[1 - \hat{p}]}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1),$$

d'où un intervalle de confiance (à 95%) pour p de la forme

$$\left[\hat{p} \pm \frac{1.96}{\sqrt{n}} \sqrt{\hat{p}[1 - \hat{p}]} \right].$$

Lois Binomiales : Inférence

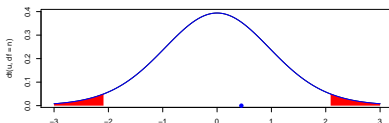
On peut faire un test de $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ (par exemple 50%). On peut utiliser le test de Student,

$$T = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}$$

qui tend, sous H_0 , vers une loi normale

```
1 > p0 = .5
2 > (T=sqrt(n)*(pn-p0)/(sqrt(p0*(1-p0))))
3 [1] 0.4472136
4 > 0.4472136<qnorm(1-alpha/2)
5 [1] TRUE
```

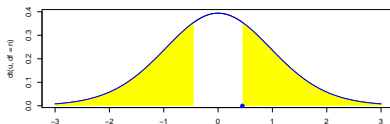
On est ici dans la région d'acceptation du test.



Lois Binomiales : Inférence

On peut aussi calculer la p -value, $\mathbb{P}(|T| > |t_{obs}|)$,

```
1 > 2*(1-pnorm(abs(T)))  
2 [1] 0.6547208
```



Lois Binomiales : Inférence

Test de Wald l'idée est d'étudier la différence entre \hat{p} et p_0 . Sous H_0 ,

$$T = n \frac{(\hat{p} - p_0)^2}{I^{-1}(p_0)} \xrightarrow{\mathcal{L}} \chi^2(1)$$

Test du rapport de vraisemblance l'idée est d'étudier la différence entre $\log \mathcal{L}(\hat{p})$ et $\log \mathcal{L}(p_0)$. Sous H_0 ,

$$T = 2 \log \left(\frac{\mathcal{L}(p_0)}{\mathcal{L}(\hat{p})} \right) \xrightarrow{\mathcal{L}} \chi^2(1)$$

Test du score l'idée est d'étudier la différence entre $\frac{\partial \log \mathcal{L}(p_0)}{\partial p}$ et 0. Sous H_0 ,

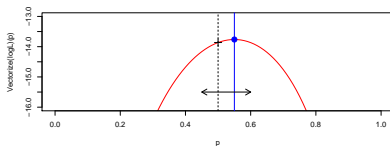
$$T = \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial \log f_{p_0}(x_i)}{\partial p} \right)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$$

Lois Binomiales : Inférence

Test de Wald différence entre \hat{p} et p_0 , test de Wald. Sous H_0 ,

$$T = n \frac{(\hat{p} - p_0)^2}{I^{-1}(p_0)} \xrightarrow{\mathcal{L}} \chi^2(1)$$

```
1 > neglogL = function(p){-sum(log(dbinom(X,1,p)))}  
2 > (IF = 1/(p0*(1-p0)/n))  
3 [1] 80  
4 > pml=optim(fn=neglogL,par=p0,method="BFGS")$par  
5 > (T=(pml-p0)^2*IF)  
6 [1] 0.199997  
7 > T<qchisq(1-alpha,df=1)  
8 [1] TRUE
```

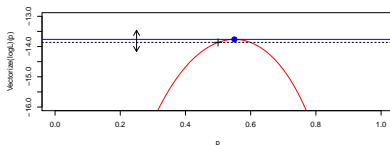


Lois Binomiales : Inférence

Test du rapport de vraisemblance l'idée est d'étudier la différence entre $\log \mathcal{L}(\hat{p})$ et $\log \mathcal{L}(p_0)$, LRT. Sous H_0 ,

$$T = 2 \log \left(\frac{\mathcal{L}(p_0)}{\mathcal{L}(\hat{p})} \right) \xrightarrow{\mathcal{L}} \chi^2(1)$$

```
1 > logL = function(p){sum(log(dbinom(X,1,p)))}  
2 > (T=2*(logL(pml)-logL(p0)))  
3 [1] 0.2003347  
4 > T<qchisq(1-alpha,df=1)  
5 [1] TRUE
```

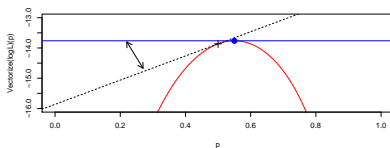


Lois Binomiales : Inférence

Test du score comparer $\frac{\partial \log \mathcal{L}(p_0)}{\partial p}$ et 0, test de Rao. Sous H_0

$$T = \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial \log f_{p_0}(x_i)}{\partial p} \right)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$$

```
1 > nx=sum(X==1)
2 > f = expression(nx*log(p)+(n-nx)*log(1-p))
3 > Df = D(f, "p")
4 > p=p0
5 > score=eval(Df)
6 > (T=score^2/IF)
7 [1] 0.2
```



Lois Multinomiales : Inférence

$\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ 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. $p_d = 1 - \mathbf{p}_{(d)}^\top \mathbf{1}$

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

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

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

Lois Multinomiales : Inférence

L'estimateur du maximum de vraisemblance est

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

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

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

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{\Omega}),$$

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

Lois Multinomiales : Inférence

Test de Pearson: $H_0 : \mathbf{p} = \mathbf{p}_0$, on utilise

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

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

On retrouvera ce test comme test d'indépendance.

Loi de Bernoulli : Application

y : indicatrice de survie d'un passager du Titanic

```
1 > loc = "http://freakonometrics.free.fr/titanic.RData"
2 > download.file(loc, "titanic.RData")
3 > load("titanic.RData")
4 > base = base[,1:7]
5 > n = nrow(base)
6 > (p = mean(base$Survived))
7 [1] 0.3838384
8 > p + qnorm(c(.025, .975))/sqrt(n)*sqrt(p*(1-p))
9 [1] 0.3519060 0.4157707
```

Si $p = \mathbb{P}(Y = 1)$, $\hat{p} = \frac{342}{891} = 38.38\%$, et

$$\mathbb{P}(p \in [35.19\%; 41.58\%]) = 95\%.$$

Loi de Bernoulli : Application

y : port d'embarquement des passagers du Titanic,
(Cherbourg, Queenstown, Southampton), $y = (1_C, 1_Q, 1_S)$

```
1 > loc = "http://freakonometrics.free.fr/titanic.RData"
2 > download.file(loc, "titanic.RData")
3 > load("titanic.RData")
4 > base = base[base$Embarked!="",]
5 > table(base$Embarked[base$Survived == 1])/sum(
    base$Survived == 1)
6
7           C           Q           S
8 0.27352941 0.08823529 0.63823529
```

$$\text{Ici } \hat{p} = \left(\frac{93}{340}, \frac{30}{340}, \frac{217}{340} \right) = (27.4\%, 8.8\%, 63.8\%)$$

