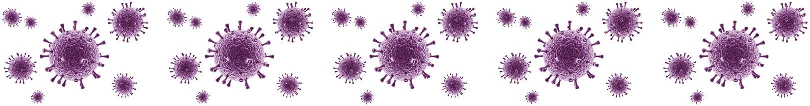
Modèles Linéaires Appliqués / Régression
Variable y catégorielle

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 $\mathbb{E}[Y] = np$.

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

Lois Binomiales & Multinomiales

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

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

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

$\boldsymbol{Y} = (Y_1, \cdots, Y_d) \sim \mathcal{M}(n, \boldsymbol{p})$ où $\boldsymbol{p} = (p_1, \cdots, p_d)$ 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: $\boldsymbol{Y} = (Y_0, Y_1) \sim \mathcal{M}(n, \boldsymbol{p})$ où $\boldsymbol{p} = (p_0, p_1)$.

Lois Binomiales : Inférence

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

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

et la log-vraisemblance est

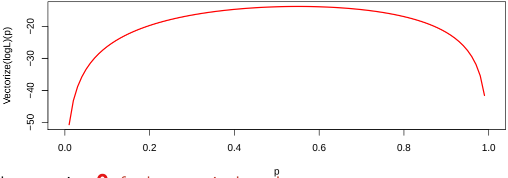
$$\log \mathcal{L}(p; \boldsymbol{y}) = \sum_{i=1}^{n} y_i \log[p] + (1-y_i)\log[1-p]$$

La condition du premier ordre est

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

Lois Binomiales : Inférence

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



Lois Binomiales : Inférence

Propriété du maximum de vraisemblance

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

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

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

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

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

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

Lois Binomiales : Inférence

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

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

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

$$\sqrt{n}\frac{\widehat{p} - p}{\sqrt{\widehat{p}[1-\widehat{p}]}} \overset{\mathcal{L}}{\rightarrow} \mathcal{N}(0,1),$$

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

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

Lois Binomiales : Inférence

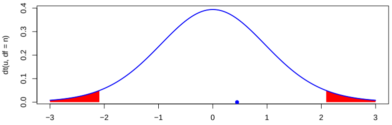
On peut faire un test de $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ (par exemple 50%). On peut utiliser le test de Student,

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

qui suit, sous H_0 une loi de Student à n degrés de liberté.

```
> p0 = .5
> (T=sqrt(n)*(pn-p0)/(sqrt(p0*(1-p0))))
[1] 0.4472136
> abs(T)<qt(1-alpha/2,df=n)
[1] TRUE
```

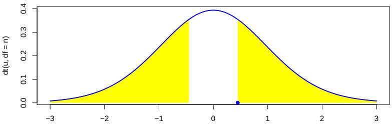
On est ici dans la région d'acceptation du test.



Lois Binomiales : Inférence

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

```
>   2*(1-pt(abs(T),df=n))
[1] 0.6595265
```



Lois Binomiales : Inférence

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

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

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

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

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

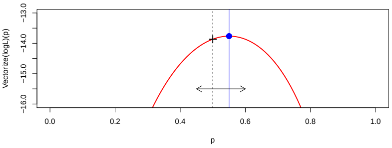
$$T = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f_{p_0}(x_i)}{\partial p}\right)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$$

Lois Binomiales : Inférence

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

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

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

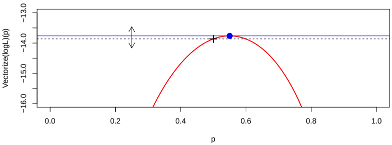


Lois Binomiales : Inférence

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

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

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

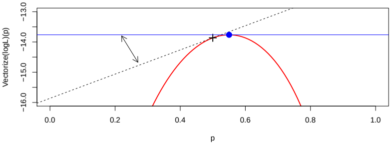


Lois Binomiales : Inférence

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

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

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



Lois Multinomiales : Inférence

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

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

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

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

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

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

Lois Multinomiales : Inférence

L'estimateur du maximum de vraisemblance est

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

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

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

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

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

Lois Multinomiales : Inférence

Test de Pearson: $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$, 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.

Loi de Bernoulli : Application

y : indicatrice de survie d'un passager du Titanic

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

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

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

Loi de Bernoulli : Application

$\boldsymbol{y}$: port d'embraquement des passagers du Titanic, (Cherbourg, Queenstown, Southampton), $\boldsymbol{y} = (\mathbf{1}_C, \mathbf{1}_Q, \mathbf{1}_S)$

```
> loc = "http://freakonometrics.free.fr/titanic.RData"
> download.file(loc, "titanic.RData")
> load("titanic.RData")
> base = base[base$Embarked!="",]
> table(base$Embarked[base$Survived == 1])/sum(
   base$Survived == 1)

         C          Q          S
0.27352941 0.08823529 0.63823529
```

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