Examen de Statistique
Master Statistique & Économétrie (M1) 2016/2017

Examen sans document, et sans calculatrice.

Pour rappels, on dira que X suit une loi Beta $\mathcal{B}(\alpha,\beta)$ si sa densité est

$$f(x)=\begin{cases}\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} & \text{pour } x\in[0,1],\\ 0 & \text{sinon.}\end{cases}$$

1 Soit $X_1,\cdots,X_5$ un échantillon i.id. de loi $\mathcal{U}([0,1])$. Quelle est la probabilité que les 4 plus petites observations soient dans l'intervalle $[0,1/2]$, et pas la plus grande ?

Regardons par simulations

```
> f=function(u) (sort(u)[4]<.5)&(sort(u)[5]>.5)
> r=function(n) f(runif(n))
> mean(Vectorize(r)(rep(5,1e6)))
[1] 0.156239
```

On obtient numériquement (presque) 5/32=0.15625. On peut aussi prouver le résultat théoriquement, par exemple en faisant un arbre, avec pour chaque noeud x_i soit $x_i\in[0,1/2)$ (pour une des branches) soit $x_i\in[1/2,1)$ (pour l'autre branche). On suppose que le premier noeud est la valeur de x_1, le second la valeur de x_2, etc, jusqu'à x_5. On va le faire sur les observations, et pas les statistiques d'ordre parce que dans ce cas les probabilités de chacune des branches sont plus simples à calculer. Avec 5 observations, on arrive à $2^5=32$ cas possibles, équiprobables (car $\mathbb{P}[X_i>1/2]=\mathbb{P}[X_i\leq 1/2]$). Or seulement 5 correspondent à la situation qui nous intéresse, donc la probabilité est 5/32. On peut aussi le faire en passant par les indicatrices $Y_i=\mathbf{1}[X_i>1/2]$, et on nous demande la probabilité que le quintuplet $\{y_1,y_2,y_3,y_4,y_5\}$ soit une permutation de $\{0,0,0,0,1\}$.

2 On considère un échantillon i.id. de taille n, tiré suivant une loi de densité

$$f_\theta(x)=\theta(1+\theta)x^{\theta-1}(1-x)\text{ pour }x\in[0,1].$$

Calculer l'estimateur de la méthode des moments de θ, noté $\widehat{\theta}$. Montrer que cet estimateur n'est pas efficace.

On pourra admettre que

$$\text{Var}(\widehat{\theta})\approx\frac{\theta(\theta+2)^2}{2n(\theta+3)}.$$

1. L'espérance de X est

$$\mathbb{E}[X]=\int_0^1 x\theta(\theta+1)x^{\theta-1}(1-x)dx=\theta(1+\theta)\left[\frac{x^{\theta+1}}{\theta+1}-\frac{x^{\theta+2}}{\theta+2}\right]_0^1=\frac{\theta}{\theta+2}.$$

de telle sorte que l'estimateur de la méthode des moments est $\widehat{\theta}$ solution de

$$\frac{\widehat{\theta}}{2+\widehat{\theta}}=\overline{x}\text{ c'est à dire }\widehat{\theta}=\frac{2\overline{x}}{1-\overline{x}}.$$

Petite remarque... 'beaucoup' on fait une erreur de calcul lors du calcul de l'espérance, et sont arrivé à $\mathbb{E}[X]=1$. Je pense qu'il aurait été donc de se dire qu'il y avait un soucis. Je laisse les plus curieux essayer d'imaginer à quoi ressemble une loi définie sur $[0,1]$, i.e. $\mathbb{P}[X\in[0,1]=1]$, vérifiant $\mathbb{E}[X]=1$...

2. Commençons par calculer l'information de Fisher :

$$\log f_\theta(x)=\log\theta+\log(1+\theta)-(\theta-1)\log x+\log(1-x)$$

soit

$$\frac{\partial}{\partial\theta}\log f_\theta(x)=\frac{1}{\theta}+\frac{1}{\theta+1}-\log x$$

de telle sorte que

$$\frac{\partial}{\partial\theta}\log f_\theta(x)=-\frac{1}{\theta^2}-\frac{1}{(\theta+1)^2}$$

et donc, si $X\sim f_\theta$

$$I_\theta=-\mathbb{E}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)=\frac{\theta^2+(\theta+1)^2}{(\theta+1)^2\theta^2}.$$

La borne de Cramér-Rao est ici

$$\frac{(\theta+1)^2\theta^2}{n(\theta^2+(\theta+1)^2)},$$

et si on regarde la différence entre $\mathrm{Var}(\widehat{\theta})$ et cette borne, on obtient

$$\frac{\theta(\theta+2)^2}{2n(\theta+3)}-\frac{(\theta+1)^2\theta^2}{n(\theta^2+(\theta+1)^2)}=\frac{\theta[3\theta^2+6\theta+4]}{2(\theta+3)}=\frac{\theta[3\theta^2+6\theta+4]}{2(\theta+3)(2\theta^2+2\theta+1)n}>0,$$

tous les termes du ratio étant positifs. Cet estimateur n'est pas efficace.

3 Considérons un échantillon $\{x_1,\cdots,x_n\}$ tiré suivant une loi Beta $\mathcal{B}(\theta,\theta)$.

1. Trouver une statistique exhaustive pour θ pour un échantillon tiré suivant une loi Beta $\mathcal{B}(\theta,\theta)$.
2. Trouver un test de niveau α de $H_0:\theta=1$ contre $H_1:\theta=2$.

1. La densité de la loi $\mathcal{B}(\theta,\theta)$ est

$$f_\theta(x)=\frac{x^{\theta-1}(1-x)^{\theta-1}}{B(\theta)}=\frac{[x(1-x)]^{\theta-1}}{B(\theta)}\text{ où }B(\theta)=\int_0^1[u(1-u)]^{\theta-1}du$$

Pour rappel, on a un statistique exhaustive $S(x)$ si la densité $f_\theta(x)$ peut s'écrire

$$f_\theta(x)=h(x)\,g(\theta,S(x))$$

Or ici, si on pose $s=x(1-x)$, on a $g(\theta,s)=s^{\theta-1}/B(\theta)$, donc d'aprs le théorème de factorisation, $S(x)=x(1-x)$ est une statistique exhaustive pour θ.

2. Le plus simple est de regarder le test du rapport de vraisemblance (Neyman-Pearson), i.e.

$$T=\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)}\text{ où }\mathcal{L}(\theta)=\prod_{i=1}^n f_\theta(x_i).$$

Or ici

$$\mathcal{L}(\theta)=\frac{1}{B(\theta)^n}\prod_{i=1}^n[x_i(1-x_i)]^{\theta-1}=\frac{1}{B(\theta)^n}\left(\prod_{i=1}^n x_i(1-x_i)\right)^{\theta-1}$$

La statistique du rapport de vraisemblance va alors s'écrire

$$T=\left(\frac{B(\theta_0)}{B(\theta_1)}\right)^n\left(\prod_{i=1}^n x_i(1-x_i)\right)^{\theta_1-\theta_0}$$

Ici, $\theta_1 - \theta_0 = 1$, donc la statistique se simplifie un peu. Et on rejettera l'hypothèse H_0 si T est "trop grande", i.e. si

$$\prod_{i=1}^{n} x_i(1-x_i) \geq c.$$

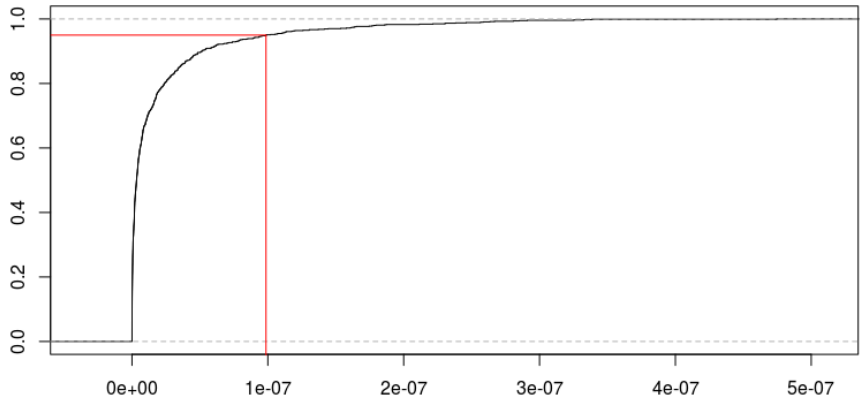
Le seuil c est ici choisit de telle sorte que

$$\mathbb{P}\left[\prod_{i=1}^{n} X_i(1-X_i) \middle| H_0\right] = \alpha$$

où le conditionnement par H_0 signifie que les variables X_i sont indépendentes, de loi $\mathcal{B}(1,1)$.

On pouvait s'arrêter là, mais on peut aussi noter que la loi $\mathcal{B}(1,1)$ est une loi uniforme sur $[0,1]$. Si on note P le produit de variables $X_i(1-X_i)$, on a besoin du quantile de cette variable pour déterminer le seuil c. Par exemple avec $n = 10$ et $\alpha = 5\%$, on utilise le code suivant

```
> n=10
> U=runif(1e4)
> M=matrix(U*(1-U),ncol=n)
> L=apply(M,1,prod)
> (q=quantile(L,.95))
[1] 9.864234e-08
> plot(ecdf(L))
```



4 Une année, on observe la naissance de 528 enfants prématurés, dont 289 garçons. Proposer une procédure de test de H_0 : "*les garçons ont plus de chances d'être prématurés que les filles*".

Formalisons un peu. Soit X_i l'indicatrice correspondant à la naissance d'un garçon : $X_i = 0$ si l'enfant i est une fille, et $X_i = 1$ si l'enfant i est un garçon. On note $N = X_1 + \cdots + X_n$. N suit ici une loi binomiale $\mathcal{B}(n,\theta)$ où θ et la probabilité d'avoir un garçon. On veut tester ici $H_0 : p = 1/2$ contre $H_1 : p > 1/2$. Le plus simple est faire un test asymptotique Gaussien, et de calculer la p-value,

$$p\text{-value} = \mathbb{P}\left[S > 284 | H_0\right] = \mathbb{P}\left[\sum_{i=1}^{n} X_i > 284 \middle| X_i \sim \mathcal{B}(1/2)\right]$$

que l'on peut réécrire

$$p\text{-value} = \mathbb{P}\left[\frac{S-np}{\sqrt{np(1-p)}} > \frac{284-np}{\sqrt{np(1-p)}} \middle| H_0\right]$$

À gauche, on a une loi normal (approchée), et à droite, on a nuériquement 2.5. La p-value est ici

$$p\text{-value} = \mathbb{P}[Z > 2.5] < 0.1\% \text{ avec } Z \sim \mathcal{N}(0,1).$$

5 On dispose d'un échantillon $\{x_1, \cdots, x_n\}$ tiré suivant un mélange

$$X = \begin{cases} U_1 \sim \mathcal{U}([0,a]) \text{ avec probabilité } \theta \\ U_2 \sim \mathcal{U}([0,b]) \text{ avec probabilité } 1-\theta \end{cases}$$

avec $\theta \in (0,1)$ et $a < b$. Soit N le nombre d'observations entre 0 et a dans l'échantillon. Quelle est la loi de N ? Quel est l'estimateur du maximum de vraisemblance de θ ?

Là encore, on peut commencer par quelques simulations de cette loi, avec $a = 1$, $b = 2$... et $\theta = .3$ parce qu'il faut bien prendre une valeur

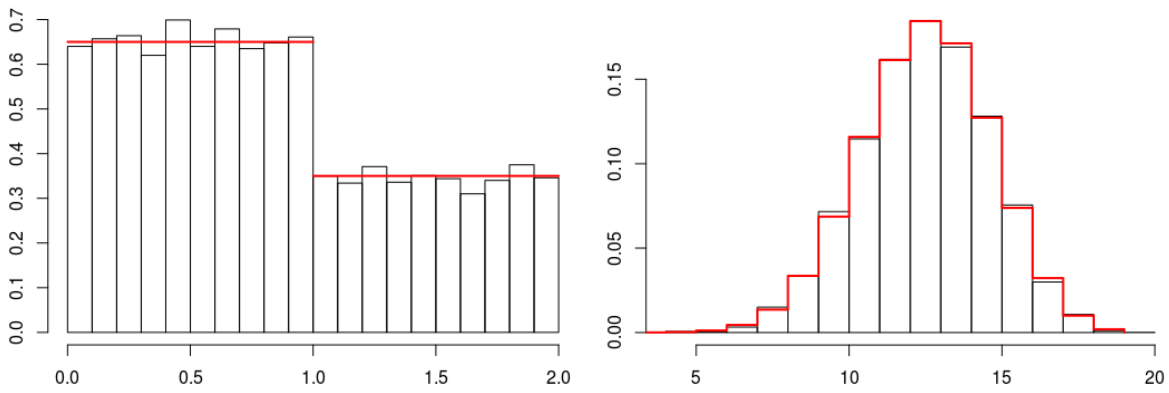
```
> a=1
> b=2
> n=1e4
> theta=.3
> T=sample(c(a,b),size=n,prob=c(theta,1-theta),replace=TRUE)
> X=runif(n,min=0,max=T)
> hist(X,prob=TRUE)
> segments(0,(1-theta)/b*(b-a)+theta,a,(1-theta)/b*(b-a)+theta,col="red",lwd=2)
> segments(a,(1-theta)/b*(b-a),b,(1-theta)/b*(b-a),col="red",lwd=2)
```

On a simulé ici un échantillon de x_i. On peut ensuite regarder la distribution des n_i en générant plusieurs échantillons.

```
> randomN=function(nb=1000,n=20){
>    N=rep(NA,nb)
>    for(i in 1:nb){
>    T=sample(c(a,b),size=n,prob=c(theta,1-theta),replace=TRUE)
>    X=runif(n,min=0,max=T)
>    N[i]=sum(X<a)}
>    return(N)
> }
```

On peut comparer l'histogramme empirique avec un loi binomiale

```
> N=randomN(1e4)
> hist(N,prob=TRUE)
> pb=theta+(1-theta)*(b-a)/b
> p=dbinom(1:20,prob=pb,size=20)
> lines(0:19,p,type="s",col="red",lwd=2)
```



Pour les calculs, on avait fait l'exo en TD. Posons $Y = \mathbf{1}(X \leq a)$. Comme les variables X_i sont indépendantes, les variables Y_i le sont également. Et comme on a des variables Y_i sont des variables distribuées suivant une loi de Bernoulli, N suit une loi binomiale. Plus précisément $N \sim \mathcal{B}(n, p = \mathbb{P}[Y = 1])$. Or ici

$$p = \mathbb{P}[X \leq a] = \underbrace{\mathbb{P}[X \leq a | X = U_1]}_{\mathbb{P}[U_1 \leq a]=1} \cdot \underbrace{\mathbb{P}[X = U_1]}_{\theta} + \underbrace{\mathbb{P}[X \leq a | X = U_2]}_{\mathbb{P}[U_2 \leq a]=\frac{a}{b}} \cdot \underbrace{\mathbb{P}[X = U_2]}_{1-\theta}$$

soit

$$p = \theta + \frac{a}{b}(1-\theta) = \frac{\theta b + (1-\theta)a}{b} = \frac{b-a}{b}\theta + \frac{1}{b}$$

qui est une fonction linéaire en θ.

Pour l'estimateur du maximum de vraisemblance de θ on peut utiliser une propriété que nous avons vu lorsqu'on reparamétrise un modèle statistique : si $\widehat{p}$ est l'estimateur du maximum de vraisemblance de p, et si $\theta = g(p)$ alors l'estimateur du maximum de vraisemblance de θ est $\widehat{\theta} = g(\widehat{p})$. Ici, on utilise le fait que l'estimation du maximum de vraisemblance de la probabilité dans un modèle binomial est la proportion. Autrement dit, $\widehat{p}$ est la proportion d'observations inférieure à a, i.e.

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq a).$$

Et à partir de là, on en déduit l'estimateur du maximum de vraisemblance de θ puisque $\widehat{\theta}$ est solution de

$$\frac{b-a}{b}\widehat{\theta} + \frac{1}{b} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq a).$$

soit

$$\widehat{\theta} = \frac{b}{b-a}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq a) - \frac{1}{b}\right).$$