TP1 - Estimation d'une proportion dans le contexte d'une question délicate et étude de l'impact de l'échantillonnage

Fabien JOSSAUD & Fanny REBIFFE

1/15/2021

Exercice 1 - Méthode de Warner

Pour cette exercice on va créer la fonction questDelicateMeth1 qui génère des simulations de réponses selon la méthode de Warner :

```
questDelicateMeth1 <- function(n, P, theta, n.sim, p.neg=0)
{P.chap <- rep(NA,n.sim)
P <- (1-p.neg)*P
for(i in 1:n.sim){
  # Choix de la question
  question <- sample(c(1,2),n,prob=c(theta,1-theta),replace=TRUE)
  # Choix du répondant
  repondant <- sample(c(1,2),n,prob=c(P,1-P),replace=TRUE)
  #
  nb.oui <- sum(question-repondant==0)
  P.chap[i] <- (nb.oui/n+theta-1)/(2*theta-1)
}
return(P.chap=P.chap)
}
```

On va ensuite faire 1000 simulations de la méthode avec n=500 et θ=0.8

```
N <- 1000
n <- 500
theta <- 0.8
P <- 0.1
set.seed(2010)
Warner1 <- questDelicateMeth1(n,P,theta,N)
```

a - Moyenne et écart-type

```
moy1 <- mean(Warner1)
sd1 <- sqrt(var(Warner1))
```

On obtient une moyenne de 0.0989133 et un écart type de 0.0331006.

b - Valeurs théoriques

Théoriquement, on devrait avoir une moyenne à 0.1 (qui correspond à π_A) et un écart-type à 0.

c - Méthode directe

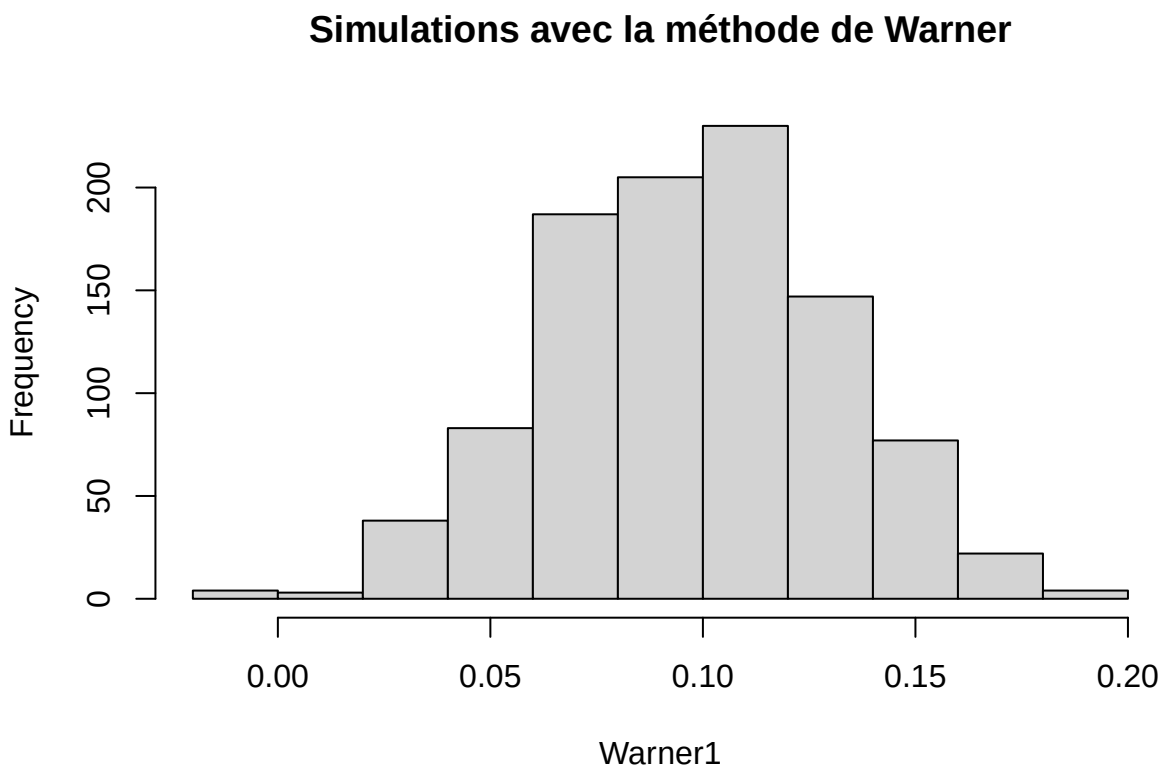
Si toutes les personnes répondent honnêtement à la question directe, le sondage correspond à une loi binomiale avec n=500 et p=0.1.

```
set.seed(2010)
simubino <- rbinom(N,n,P)/n
varbino1 <- var(simubino)
varWarner1 <- var(Warner1)
```

La variance de la méthode directe est de 1.8383305×10^{-4} et celle que la méthode de Warner est 0.0010956. Il y a une différence entre les 2 variances. En effet, même si c'est minime, la méthode directe est plus précise que la méthode de Warner à condition que la population réponde honnêtement.

d - Histogramme

```
hist(Warner1,main="Simulations avec la méthode de Warner")
```



e - Commentaires

On remarque que la méthode de Warner se rapproche assez bien de ce qu'on espère obtenir c'est à dire une estimation $\hat{\pi}_{AW}$ proche de π_A. Cependant, la méthode de Warner est moins efficace que la méthode directe à la condition que les gens honnêtement au sondage.

Exercice 2 - Méthode de la question complémentaire innocente

Comme précédemment, on va créer la fonction questDelicateMeth2 qui génère des simulations de réponses selon la méthode de la question complémentaire :

```
questDelicateMeth2 <- function(n, P, theta, alpha, n.sim, p.neg=0)
{P.chap <- rep(NA,n.sim)
P <- (1-p.neg)*P
for(i in 1:n.sim){
  # Choix de la question
  question <- sort(sample(c(1,2),n,prob=c(theta,1-theta),replace=TRUE))
  # La question étant choisie, à qui s'adresse-t-elle ?
  nb.1 <- sum(question==1)
  nb.oui <- sum(sample(c(1,2),nb.1,prob=c(P,1-P),replace=TRUE)==1)+
    sum(sample(c(1,2),n-nb.1,prob=c(alpha,1-alpha),replace=TRUE)==1)
  P.chap[i] <- (nb.oui/n-(1-theta)*alpha)/(theta)
}
return(P.chap=P.chap)
}
```

On va ensuite faire 1000 simulations de la méthode avec n=500, θ=0.8 et α=0.217.

```
N <- 1000
n <- 500
theta <- 0.8
alpha <- 0.217
P <- 0.1
set.seed(2010)
Meth2 <- questDelicateMeth2(n,P,theta,alpha,N)
```

a - Moyenne et écart-type

```
moy2 <- mean(Meth2)
sd2 <- sqrt(var(Meth2))
```

On obtient une moyenne de 0.09867 et un écart type de 0.0180145.

b - Valeurs théoriques

On a toujours 0.10 et 0 en valeurs théoriques.

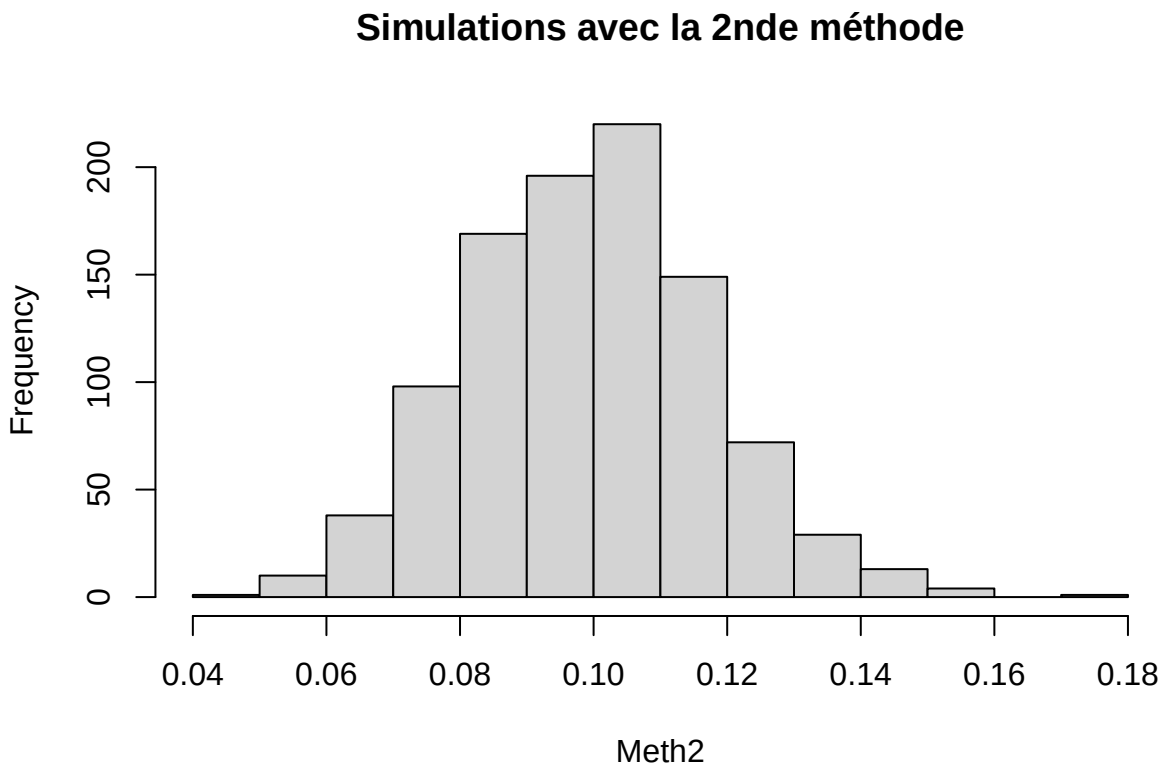
c - Comparaison avec la méthode directe

```
varMeth2 <- var(Meth2)
```

La variance de la méthode directe est de 1.8383305×10^{-4} et celle que la méthode de la question complémentaire est 3.2452312×10^{-4}. On remarque toujours une différence entre les 2 méthodes même si elle est encore plus infime que pour la méthode de Warner.

d - Histogramme

```
hist(Meth2,main="Simulations avec la 2nde méthode")
```



e - Commentaires

On remarque que, dans notre cas, la variance est significativement plus faible, ce qui indique que la seconde méthode est plus reproductible que pour Warner mais la moyenne s'éloigne de la valeur espérée, la seconde méthode est biaisée.

Exercice 3 - Comparaison des méthodes si un certain pourcentage des répondants nient

0 %

```
eqmMeth2 <- var(Meth2) + (mean(Meth2)-P)^2
eqmWarner0 <- var(Warner1) + (mean(Warner1)-P)^2
```

Ici, l'EQM de la méthode de Warner est 0.0010968 et 3.2629202×10^{-4} pour la seconde méthode. Dans le cas où personnes ne nient pas, la Méthode 2 possède la plus petite EQM.

10 %

```
set.seed(2010)
Warner10 <- questDelicateMeth1(n,P,theta,N,0.1)
eqmWarner10 <- var(Warner10) + (mean(Warner10)-P)^2

set.seed(2010)
Meth210 <- questDelicateMeth2(n,P,theta,alpha,N,0.1)
eqmMeth210 <- var(Meth210) + (mean(Meth210)-P)^2
```

Ici, l'EQM de la méthode de Warner est 0.0011765 et 4.2036975×10^{-4} pour la seconde méthode.

20 %

```
set.seed(2010)
Warner20 <- questDelicateMeth1(n,P,theta,N,0.2)

eqmWarner20 <- var(Warner20) + (mean(Warner20)-P)^2

set.seed(2010)
Meth220 <- questDelicateMeth2(n,P,theta,alpha,N,0.2)
eqmMeth220 <- var(Meth220) + (mean(Meth220)-P)^2
```

Ici, l'EQM de la méthode de Warner est 0.0014692 et 7.1283164×10^{-4} pour la seconde méthode.

30 %

```
set.seed(2010)
Warner30 <- questDelicateMeth1(n,P,theta,N,0.3)

eqmWarner30 <- var(Warner30) + (mean(Warner30)-P)^2

set.seed(2010)
Meth230 <- questDelicateMeth2(n,P,theta,alpha,N,0.3)
eqmMeth230 <- var(Meth230) + (mean(Meth230)-P)^2
```

Ici, l'EQM de la méthode de Warner est 0.0019666 et 0.0012038 pour la seconde méthode.

40 %

```
set.seed(2010)
Warner40 <- questDelicateMeth1(n,P,theta,N,0.4)

eqmWarner40 <- var(Warner40) + (mean(Warner40)-P)^2

set.seed(2010)
Meth240 <- questDelicateMeth2(n,P,theta,alpha,N,0.4)
eqmMeth240 <- var(Meth240) + (mean(Meth240)-P)^2
```

Ici, l'EQM de la méthode de Warner est 0.0026814 et 0.0019115 pour la seconde méthode.

On remarque que, quelque soit le pourcentages de négation, la 2è méthode reste la plus fiable (ce qui est logique car, comme indiqué page 3, la variance de la $2^{ème}$ méthode est plus faible à partir du moment ou θ>1/3).

Exercice 4 - Etude des paramètres statistiques empiriques et théoriques

a - Exemple d'une population de petite taille

```
pop <- c(3, 6, 24, 27, 30, 36, 51, 57)
N <- length(pop)
yu <- mean(pop)
S <- sd(pop)
```

L'espérance de la population est $\bar{y}_{\mathcal{U}}$=29.25 et son écart-type S=19.0918831.

b - Echantillonnage

```
options(OutDec =",")
n<-3
(n.ech<-choose(N,n))
```

```
## [1] 56
```

```
ech.all<-cbind(choix<-t(matrix(combn(1:N,n),n,n.ech)),(matrix(pop[choix],n.ech,n)))
rownames(ech.all)<-seq(1:nrow(ech.all))
colnames(ech.all)<-c(paste("i",1:n,sep =""),paste("pop[i",1:n ,"]", sep =""))
head(ech.all)
```

```
##   i1 i2 i3 pop[i1] pop[i2] pop[i3]
## 1  1  2  3       3       6      24
## 2  1  2  4       3       6      27
## 3  1  2  5       3       6      30
## 4  1  2  6       3       6      36
## 5  1  2  7       3       6      51
## 6  1  2  8       3       6      57
```

c - Calcul des paramètres statistiques

```
ech.all<-as.data.frame(ech.all)

ech.all$ybar<-apply(ech.all[,4:6],1,mean)
ech.all$s<-apply(ech.all[,4:6],1,sd)
ech.all$b.inf<-mapply(function(i,j){i-1.96*sqrt((1-3/8)*j^2/3)},ech.all$ybar,ech.all$s)
ech.all$b.sup<-mapply(function(i,j){i+1.96*sqrt((1-3/8)*j^2/3)},ech.all$ybar,ech.all$s)
ech.all$incl<-mapply(function(i,j){if (i<yu & j>yu) {1}else{0}},ech.all$b.inf,ech.all$b.sup)
attach(ech.all)

head(ech.all[,-(1:6)])
```

```
##   ybar        s      b.inf    b.sup incl
## 1   11 11,35782  0,8391437 21,16086    0
## 2   12 13,07670  0,3014103 23,69859    0
## 3   13 14,79865 -0,2390710 26,23907    0
## 4   15 18,24829 -1,3251646 31,32516    1
## 5   20 26,88866 -4,0549579 44,05496    1
## 6   22 30,34798 -5,1497145 49,14971    1
```

d - Vérification du lien entre paramètres empiriques sur l'ensemble des échantillons possibles et paramètres théoriques

```
muybar <- mean(ybar)
muybar == yu
```

```
## [1] TRUE
```

On a bien $\mu_{\bar{y}} = \bar{y}_{\mathcal{U}}$.

```
sigmaybarth <- sqrt(1-n/N)*(S/sqrt(n))
sigmaybar <- sd(ybar)*sqrt((length(ybar)-1)/length(ybar))
sigmaybar
```

```
## [1] 8,714213
```

```
sigmaybarth
```

```
## [1] 8,714213
```

On a bien $\sigma_{\bar{y}} = \sqrt{1-f}\frac{S}{\sqrt{n}}$.

e - Evaluation de l'estimation de l'écart-type

```
mean(s)-S
```

```
## [1] -1,186234
```

Le biais n'est donc pas nul.

f - Evaluation de l'estimation de l'espérance

```
erreur_2<-unlist(lapply(ech.all$ybar,function(i){if (abs(i-yu)>2) {1}else{0}}))
proba_2<-sum(erreur_2)/length(erreur_2)

erreur_5<-unlist(lapply(ech.all$ybar,function(i){if (abs(i-yu)>5) {1}else{0}}))
proba_5<-sum(erreur_5)/length(erreur_5)

erreur_25<-unlist(lapply(ech.all$ybar,function(i){if (abs(i-yu)>yu/4) {1}else{0}}))
proba_25<-sum(erreur_25)/length(erreur_25)
```

La probabilité de commettre une erreur de plus de : - 2 unités est de 0,75 - 5 unités est de 0,625 - 25% est de 0,4642857.

g - Evaluation de l'estimation de l'écart-type

```
erreur_20<-unlist(lapply(ech.all$s,function(i){if (abs(i-S)>S/5) {1}else{0}}))
proba_20<-sum(erreur_20)/length(erreur_20)
```

La probabilité de commettre une erreur de plus de 20% est de 0,7142857.

h - Niveau de confiance de l'intervalle de confiance d'une largeur de 4 $\hat{\sigma}_{\bar{y}}$

```
correct_ICh<-unlist(mapply(function(i,j){
  if ((i-2*sqrt((1-3/8)*j^2/3))<yu & yu<(i+2*sqrt((1-3/8)*j^2/3)))
    {1}else{0}},ech.all$ybar,ech.all$s))
proba_ICh<-sum(correct_ICh)/length(correct_ICh)
```

Le niveau de confiance de cet intervalle est 89%.

i - Niveau de confiance de l'intervalle de confiance d'une largeur de 6$\hat{\sigma}_{\bar{y}}$

```
correct_ICi<-unlist(mapply(function(i,j){
  if ((i-3*sqrt((1-3/8)*j^2/3))<yu & yu<(i+3*sqrt((1-3/8)*j^2/3)))
    {1}else{0}},ech.all$ybar,ech.all$s))
proba_ICi<-sum(correct_ICi)/length(correct_ICi)
```

Le niveau de confiance de cet intervalle est 96%.

j - Niveau de confiance de l'intervalle de confiance d'une largeur de 4 $\sigma_{\bar{y}}$

```
sigma_y_bar<-sqrt(var(ech.all$ybar))*(length(ech.all$ybar)-1)/length(ech.all$ybar)
correct_ICj<-unlist(lapply(ech.all$ybar,function(i){
  if ((i-2*sigma_y_bar)<yu & yu<(i+2*sigma_y_bar))
    {1}else{0}}))
proba_ICj<-sum(correct_ICj)/length(correct_ICj)
```

Le niveau de confiance de cet intervalle est 96%.

Exercice 5 - Etude de l'impact de l'échantillonnage sur les résutats statistiques

a - Evaluation de l'estimation de l'espérance

```
# Descriptif de population
# Valeurs
y <- c(36,36.5,37,37.5,38,38.5,39,39.5,40,40.5,41,
       41.5,42,42.5,43,43.5,44,44.5,45,45.5,46,46.5)
# Probabilites
e<-c(0.02,0.03,0.04,0.06,0.10,0.12,0.13,0.10,0.07,0.05,0.05,
     0.04,0.04,0.03,0.03,0.02,0.02,0.01,0.01,0.01,0.01,0.01)
# Echantillonnage
n<-5
set.seed(2010)
echt<-matrix(unlist(lapply(1:1000,function(i){sample(y,n,replace=TRUE,prob=e)})),ncol=5,byrow=TRUE)
```

```
y_u<-sum(unlist(mapply(function(i,j){i*j},y,e)))
y_u
```

```
## [1] 39,79
```

```
y_bar<-apply(echt, 1, mean)

mean(y_bar)
```

```
## [1] 39,798
```

Le biais entre $\mu_{\bar{y}}$ et $\bar{y}_{\mathcal{U}}$ est faible (0,008).

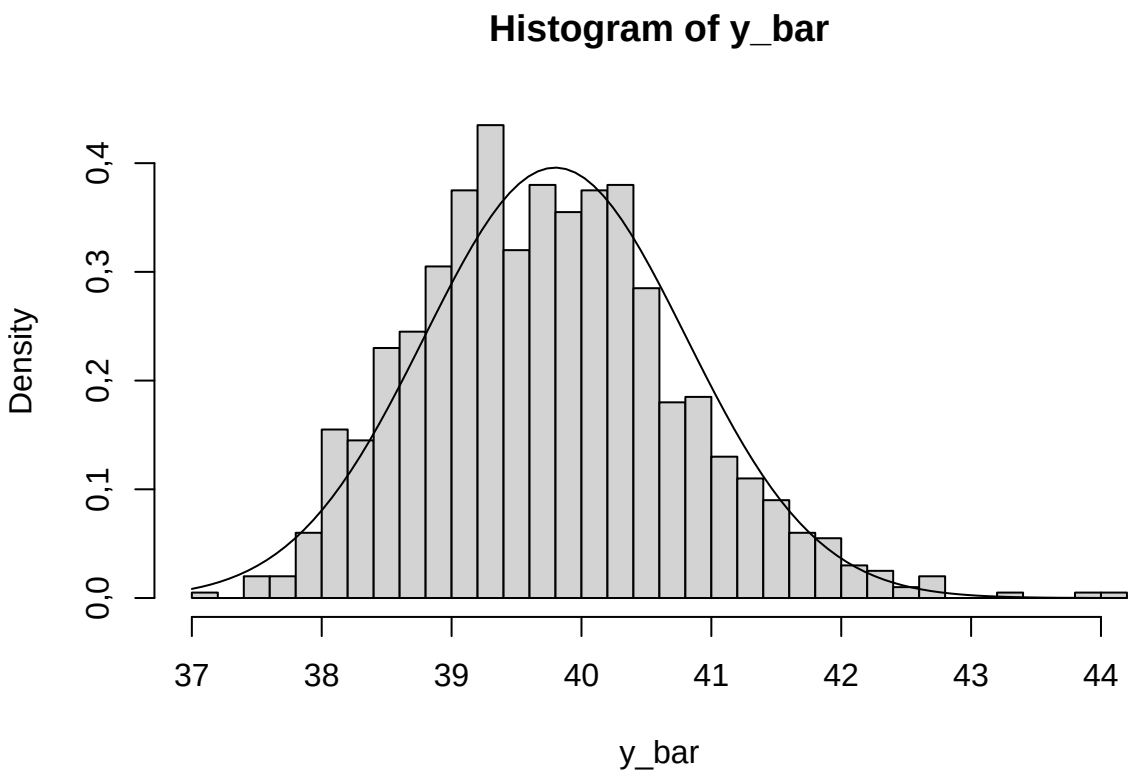
b - Evaluation de l'estimation de la variance

```
sigma2<-sum(unlist(mapply(function(i,j){(i-y_u)^2*j},y,e)))
var_y_th<-sigma2/length(y)
var_y_emp<-var(y_bar)*(length(y_bar)-1)/length(y_bar)
```

La variance est de 1,015316. Elle devrait théoriquement être $\frac{\sigma^2}{n} = 0{,}2307227$.

c - Distribution de $\bar{y}$

```
hist(y_bar,breaks=35,freq=F)
curve(dnorm(x,mean(y_bar),sqrt(var_y_emp)),add=TRUE)
```



d - Probabilité que $\bar{y}$ soit éloigné de plus de 1.96 écart-type de la moyenne

```
out<-lapply(y_bar,function(i){
  if (mean(y_bar)-1.96*var_y_emp<i & i<mean(y_bar)+1.96*var_y_emp)
    {0}else{1}})
```

La probabilité que $\bar{y}$ soit éloigné de plus de 1.96 écart-type de la moyenne est de 4%.

e - Intervale de confiance à 95%

```
n<-dim(echt)[2]
# variance empirique corrigée
s_prime<-apply(echt, 1, sd)
# Variance empirique
s<-apply(echt, 1, function(i){sqrt(var(i)*(n-1)/n)})
```

```
inside<-mapply(function(i,j){
  if (i-1.96*j/sqrt(dim(echt)[2])<y_u & y_u< i+1.96*j/sqrt(dim(echt)[2]))
    {1}else{0}},y_bar,s )
```

La probabilité que l'intervalle $\bar{y} \pm 1.96s/\sqrt{n}$ inclu $\bar{y}_{\mathcal{U}}$ est de 81%.

f - Comparaison de s^2 et S^2

Les résultats de la question e semblent indiquer que s^2 est un estimateur biaisé de S^2.

En effet, on obtient un biais de 0,95892.

Remarque : ce biais n'est plus que de 0,070325 si on prend la variance empirique corrigée.

g - Autre intervalle de confiance à 95%

```
alpha<-0.05

inside2<-lapply(s,
                function(j){
                  if ((n-1)*j^2/qchisq(1-alpha/2,n-1)<sigma2 & sigma2< (n-1)*j^2/qchisq(1-(1-alpha/2),n
                    {1}else{0}})
```

Le niveau de l'intervalle

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}} < S^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}$$

est de 94%.