Inferential Statistics Asthenment

Costalonga Andrea 2019164

Introduction

I went through the first and the third asthenment, here reported in this order. This report has been created using r markdown and latex.

First Assignment

First of all I've defined some constants and a function to make my code clear and understandable. I've also included some libraries that I've used.

```
library(plotrix)
library(rmarkdown)
set.seed(29)

#tau function
tau <- function(theta){
    out <- theta/(1-theta)
    out2 <- log(out)
    return(out2)
}

theta0 = 2/3 ##Neg bin success probability
r = 10 ##Neg bin size param
n <- c(10,50) ## Sample sizes
alpha = 0.05 ## Wald c.i. .95
```

In this segment I've created 2*N matrix with N = 10^5 to collect some results to simulate the distribution of tau hat. Every $\hat{\tau}$ is obtained evaluating $\hat{\theta}$ with the function tau. $\hat{\theta}$ is defined as

$$\hat{\theta} = \frac{r * n}{r * n + < \mathbf{y}, \mathbf{1} >}$$

where r is the number of odds of the negative binomial, n is the size of the r. ve. and $\mathbf{y}$ is the observed sample obtained from the r. ve.. This results comes as the maximum of the likelihood function.

```
N = 1e5
sim.N.tau <-matrix(NA, nrow = N, ncol = 2)
for(i in 1:N){
    xnegbin10 <- rnbinom(n=10,size=10,prob=theta0)
    xnegbin50 <- rnbinom(n=50,size=10,prob=theta0)
    sim.N.tau[i,1] <- tau(r*n[1]/(r*n[1]+sum(xnegbin10)))
    sim.N.tau[i,2] <- tau(r*n[2]/(r*n[2]+sum(xnegbin50)))
}
```

In the next cell I've calculated what would have been the true value of mean and variance of the distribution of $\hat{\tau}$. In order to do so I've calculated the variance of $\hat{\theta}$ with Cramer-Rao's theorem, knowing that the MLE is asymptotically efficient. In formulas:

$$Var(\hat{\theta}) = I_{\mathrm{n}}(\theta)^{-1}$$

where $In(\theta)$ is the Fisher information

$$I_{\mathrm{n}}(\theta) = n * I(\theta)$$

knowing that our r. ve. is built from Y1, …., Yn iid variables.

$$I(\theta) = \frac{\theta^2 * (1-\theta)}{r * n}$$

obtained from the definition of Fisher information After that I've used the delta method to obtain the variance and the mean of the gaussian curve that represent our ideal distribution: $\hat{\tau}$'s variance:

$$var(\hat{\tau}) = (\frac{dtau(\theta)}{d\theta})^2 * Var(\hat{\theta})$$

$\hat{\tau}$'s mean:

$$E[\hat{\tau}] = tau(\theta)$$

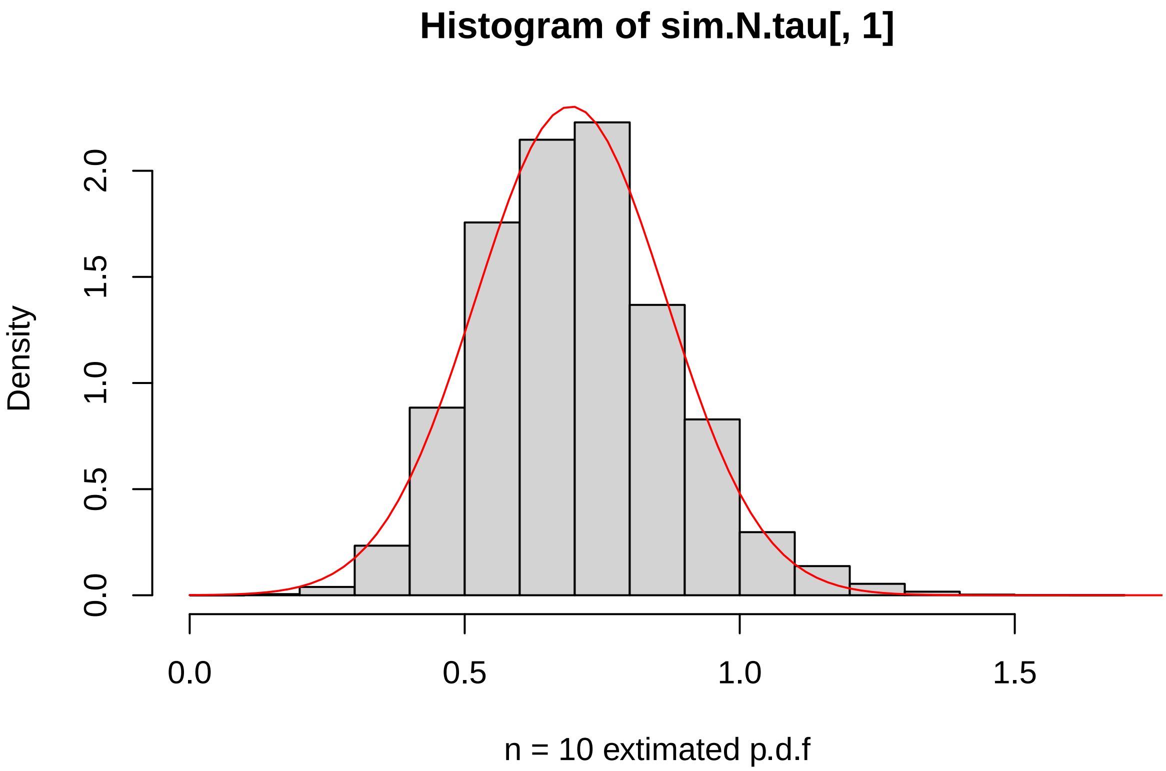
Evaluating variance and mean in θ_0 bring us to the real distribution of $\hat{\tau}$

```
true_mean = (tau(theta0))
true_var10_o <- (theta0^2*(1-theta0))/(r*n[1])
true_var10 <- true_var10_o*(1/(theta0*(1-theta0)))^2
true_var50_o <- (theta0^2*(1-theta0))/(r*n[2])
true_var50 <- true_var50_o*(1/(theta0*(1-theta0)))^2
```

I've made a couple of plot to see if our theoretical results are supported by our simulations.

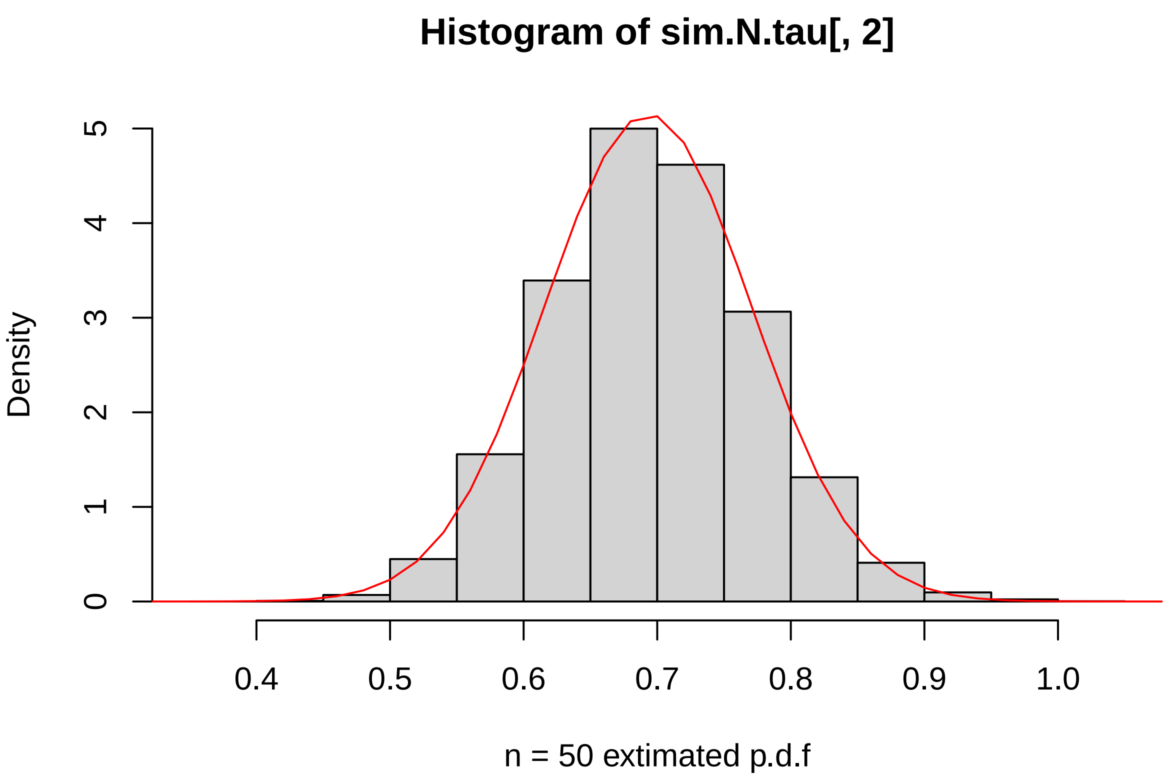
Plot for n = 10:

```
hist(sim.N.tau[,1], freq = FALSE, breaks=20, xlab = "n = 10 extimated p.d.f")
plot(function(x) dnorm(x, mean = true_mean, sd=sqrt(true_var10)),xlim=c(0,2),add=TRUE,yl
```



Plot for n = 50:

```
hist(sim.N.tau[,2], freq = FALSE, breaks=20, xlab = "n = 50 extimated p.d.f")
plot(function(x) dnorm(x, mean = true_mean, sd=sqrt(true_var50)),xlim=c(0,2),add=TRUE,yl
```



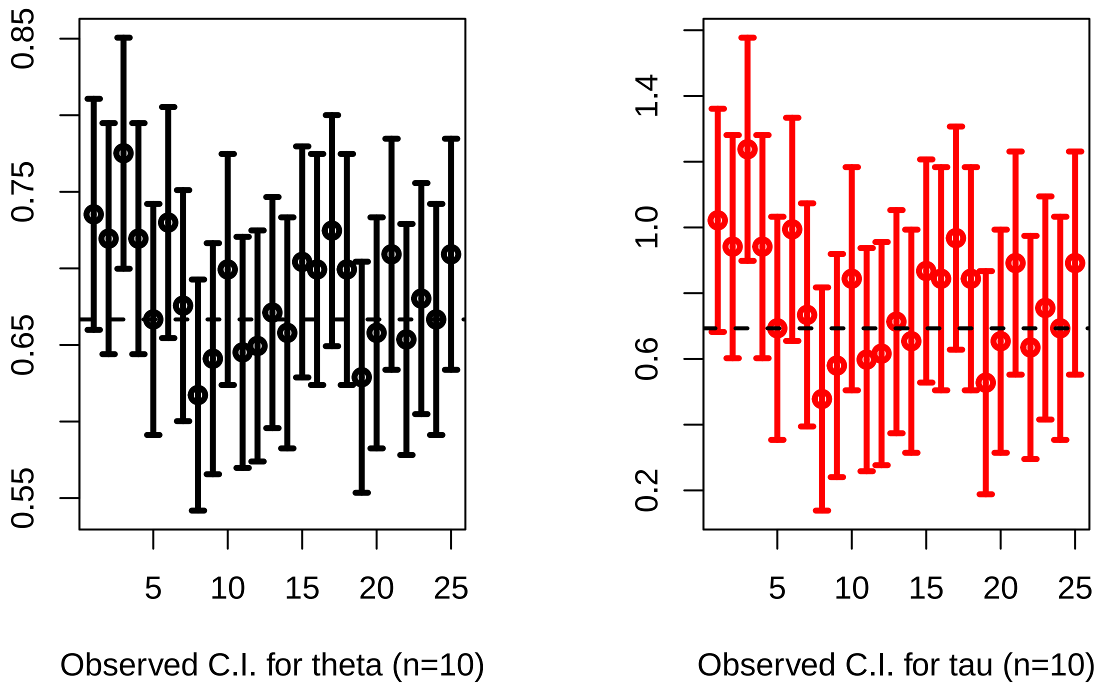
We can observe that the extimated distribution is coherent with our data.

In the following lines I've made a simulation to study the coverage probability of the .95 Wald confidene interval for τ and θ

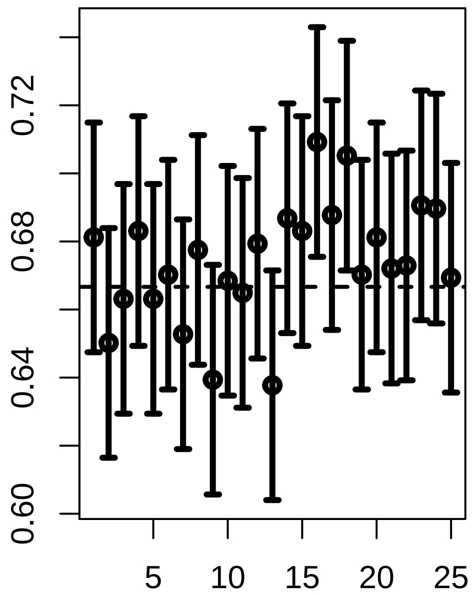
```
sim.N.CI_the <-matrix(NA, nrow = N, ncol = 4)
sim.N.CI_tau <-matrix(NA, nrow = N, ncol = 4)
for(i in 1:N){
    xnegbin10 <- rnbinom(n=n[1],size=r,prob=theta0)
    xnegbin50 <- rnbinom(n=n[2],size=r,prob=theta0)
    sim.N.CI_the[i,1:2] <- r*n[1]/(r*n[1]+sum(xnegbin10))
    sim.N.CI_the[i,1:2] <- sim.N.CI_the[i,1:2] + c(-1,1)*qnorm(p = alpha/2,
         lower.tail =FALSE)*sqrt(true_var10_o)
    sim.N.CI_the[i,3:4] <- r*n[2]/(r*n[2]+sum(xnegbin50))
    sim.N.CI_the[i,3:4] <- sim.N.CI_the[i,3:4] + c(-1,1)*qnorm(p = alpha/2,
         lower.tail =FALSE)*sqrt(true_var50_o)
    sim.N.CI_tau[i,1:2] <- tau(r*n[1]/(r*n[1]+sum(xnegbin10)))
    sim.N.CI_tau[i,1:2] <- sim.N.CI_tau[i,1:2] + c(-1,1)*qnorm(p = alpha/2,
        lower.tail =FALSE)*sqrt(true_var10)
    sim.N.CI_tau[i,3:4] <- tau(r*n[2]/(r*n[2]+sum(xnegbin50)))
    sim.N.CI_tau[i,3:4] <- sim.N.CI_tau[i,3:4] + c(-1,1)*qnorm(p = alpha/2,
        lower.tail =FALSE)*sqrt(true_var50)
}
theta10.inside <-apply(sim.N.CI_the[,1:2], MARGIN = 1,
    function(x)ifelse(theta0>=x[1]&theta0<=x[2],1,0))
theta50.inside <-apply(sim.N.CI_the[,3:4], MARGIN = 1,
    function(x)ifelse(theta0>=x[1]&theta0<=x[2],1,0))
tau10.inside <-apply(sim.N.CI_tau[,1:2], MARGIN = 1,
    function(x)ifelse(true_mean>=x[1]&true_mean<=x[2],1,0))
tau50.inside <-apply(sim.N.CI_tau[,3:4], MARGIN = 1,
    function(x)ifelse(true_mean>=x[1]&true_mean<=x[2],1,0))
```

In the plots below I've reported the first 25 confidence intervals for θ and τ

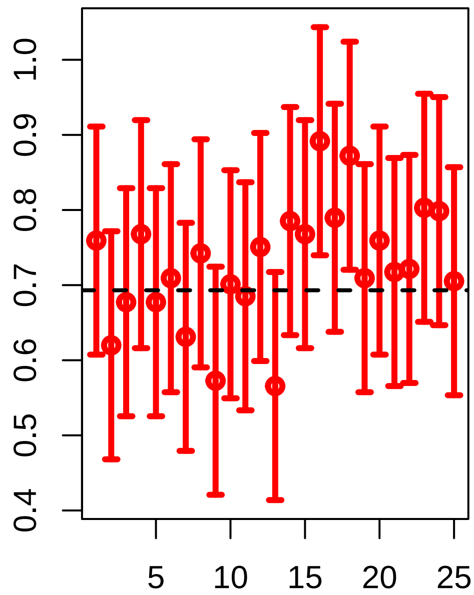
```r
par(mfrow =c(1,2))
plotCI(x= 1:25, y =apply(sim.N.CI_the[1:25,1:2], 1, mean),
    li = sim.N.CI_the[1:25,1],ui = sim.N.CI_the[1:25,2],
    xlab="Observed C.I. for theta (n=10)",ylab=NA, lwd=3)
    abline(h = theta0, lwd=2, lty=2)
plotCI(x= 1:25, y =apply(sim.N.CI_tau[1:25,1:2], 1, mean),
    li = sim.N.CI_tau[1:25,1],ui = sim.N.CI_tau[1:25,2],
    xlab="Observed C.I. for tau (n=10)",ylab=NA, lwd=3, col="red")
    abline(h = tau(theta0), lwd=2, lty=2)
```




```r
plotCI(x= 1:25, y =apply(sim.N.CI_the[1:25,3:4], 1, mean),
    li = sim.N.CI_the[1:25,3],ui = sim.N.CI_the[1:25,4],
    xlab="Observed C.I. for theta (n=50)",ylab=NA, lwd=3)
    abline(h = theta0, lwd=2, lty=2)
plotCI(x= 1:25, y =apply(sim.N.CI_tau[1:25,3:4], 1, mean),
    li = sim.N.CI_tau[1:25,3],ui = sim.N.CI_tau[1:25,4],
    xlab="Observed C.I. for tau (n=50)",ylab=NA, lwd=3, col="red")
    abline(h = tau(theta0), lwd=2, lty=2)
```



Observed C.I. for theta (n=50)



Observed C.I. for tau (n=50)

Here's some results

```
mean(theta10.inside) #Confidence probability for theta n=10
```

```
## [1] 0.95415
```

```
mean(theta50.inside)  #Confidence probability for theta n=50
```

```
## [1] 0.94976
```

```
mean(tau10.inside)  #Confidence probability for tau n=10
```

```
## [1] 0.94768
```

```
mean(tau50.inside)  #Confidence probability for tau n=50
```

```
## [1] 0.94857
```

We can notice that in every scenario the confidence probability close to .95 (even better some times). It's easy to state that our confidence interval for $\hat{\theta}$ and $\hat{\tau}$ have the same behaviour.

Third Assignment

I've defined some recurrent values and settings below:

```
set.seed(29)
n = c(10,15,10,15) #Number of elements for every col
var = c(1,1,1,1) #Variance for every col
mu = c(0,0,0,0) #Mean for every col
k=4 #Number of cols
N = 1e3 #Number of iteration for simulation
```

I've defined a matrix with dimension 2*N to store the value of F and its relative probability. This data is going to be used in a following plot.

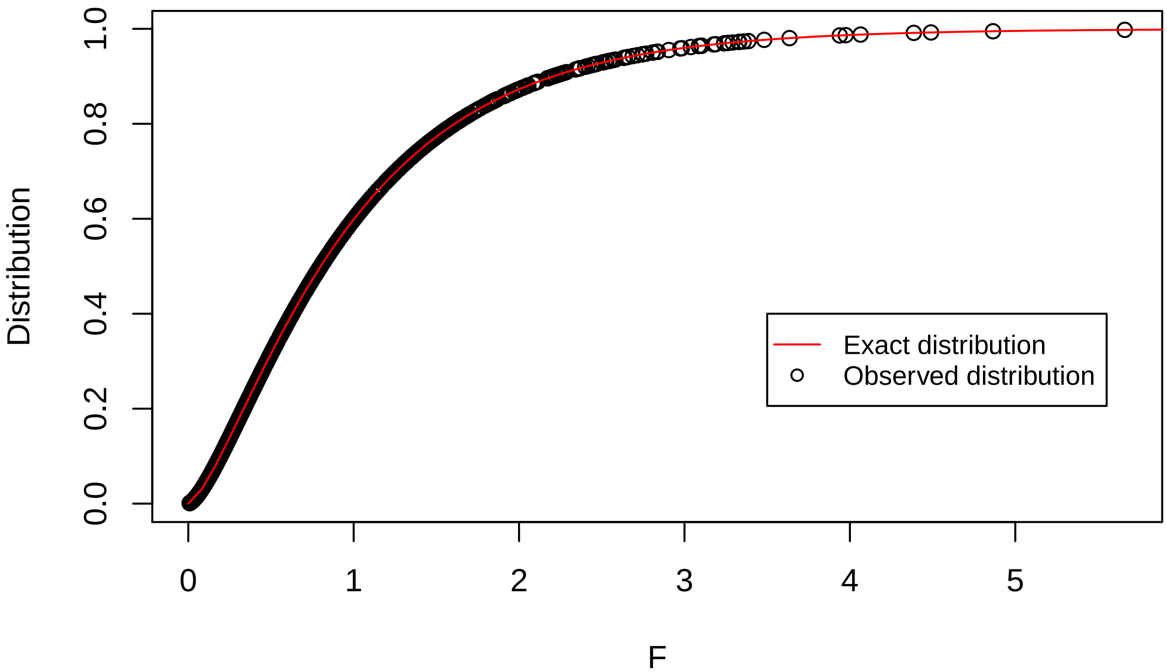
```
sim.N.F <-matrix(NA, nrow = N, ncol = 2)
for(l in 1:N){
    for (i in 1:k){
        y<-rnorm(sum(n), mean = mu[i], sd = sqrt(var[i]))
    }
    estMuj <- c(sum(y[1:n[1]])/n[1], sum(y[n[1]+1:n[2]])/n[2],
        sum(y[n[2]+1:n[3]])/n[3], sum(y[n[3]+1:n[4]])/n[4])
    estMu <- sum(y)/sum(n)
    SSR <- 0
    SSE <- 0
    for (i in 1:k){
        offset <- sum(n[1:(i-1)])
        for(j in 1:n[i]){
            SSR = SSR + (estMuj[i]-estMu)^2
            SSE = SSE + (y[offset+j]-estMuj[i])^2
        }
    }
    sim.N.F[l,1] <- (SSR/(k-1))/(SSE/(sum(n)-k)) #value of F observed (Fobs)
    sim.N.F[l,2] <- pf(sim.N.F[l,1],k-1,sum(n)-k) # P(Fk-1,n-k >=Fobs)
}
```

From theory we know that

$$\frac{\frac{SSR}{k-1}}{\frac{SSE}{n-k}} \sim \frac{\chi_{\text{k-1}}{}^2}{\chi_{\text{n-k}}{}^2}$$

This new r.v. should follow a F distribution with parameters k-1 and n-k. In the plot below there's a comparison between our simulated data and our real distribution. As we can see our observed samples are coherent with the real distribution.

```
plot(sim.N.F,xlab="F",ylab="Distribution")
plot(function(x) pf(x, df1=k-1, df2=sum(n)-k),xlim=c(0,8),add=TRUE, col='red')
legend(3.5, 0.4, legend=c("Exact distribution", "Observed distribution"),
       col=c("red", "black"), lwd=1,lty=c(1,NA), pch=c(NA,1), cex=0.8)
```



Distribution
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
F
Exact distribution
Observed distribution