Entrega final

Clara Albert

Abril 2021

Exercici 1

Apartat a

Proveu analíticament l'esperança i variància

Per provar $E(X_n) = n(p-q)$, partim de que les variables $Y_1, Y_2, \ldots, Y_n$ són variables aleatòries i independents i idènticament distribuïdes. Per tant, $E(X_n) = E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E\,(Y_i)$, on $E(Y) = (1 \cdot p + (-1) \cdot q)$ (calculada a partir de que Y és una variable discreta i per tant la seva esperança és la suma de probabilitats multiplicada pel valor que pren).

Per tant $E(X_n) = \sum_{i=1}^{n}(\,1 \cdot p + (-1) \cdot q) = sum_{i=1}^{n}(p-q) = n(p-q)$

De manera anàloga, per provar $Var(X_n) = 4npq$, partim de que les variables són aleatòries i iid i per tant $Var(X_n) = \sum_{i=1}^{n} V\,ar(Y_i)$, on $Var(Y_i) = E(Y_i^2) - [E(Y_i)]^2$

Calcularem $Var(Y_i)$ per parts:

$E(Y_i^2) = (-1)^2 \cdot q + 1^2 \cdot p = q + p = 1$

$E(Y_i)^2 = (p-q)^2 = p^2 - 2pq + q^2 = p(1-q) - 2pq + q(1-p)$
$= p - pq - 2pq + q - pq = p + q - 4pq = 1 - 4pq$

Una vegada tenim les dues esperances calculades, tenim que: $Vat(X_n) = \sum_{i=1}^{n} 1 - (1 - 4pq) = \sum_{i=1}^{n} 4\,pq = 4npq$

Distribució asimptòtica del procés segons *Teorema Central del Límit*

El teorema del límit central indica que, en condicions molt generals, si tenim una variable que es la suma de n variables aleatòries independents, amb mitjana coneguda i variància no nul·la però finita, llavors aquesta funció de distribució s'aproxima bé a una distribució normal.

En el nostre cas, $X_n = \sum_{i=1}^{n} Y_i$, per tant tenim la variable suma. A més a més, sabem que les $Y_1, \ldots Y_n$ són iid. Tal i com hem demostrat abans, $E(X) = n(p-q)$ i $Var(X) = 4npq$. Per tant, la distribució assimptòtica de X segons el teorema és:

$X_n \sim N(\,n(p-q)\,,\ 4npq\,)$

Apartat b

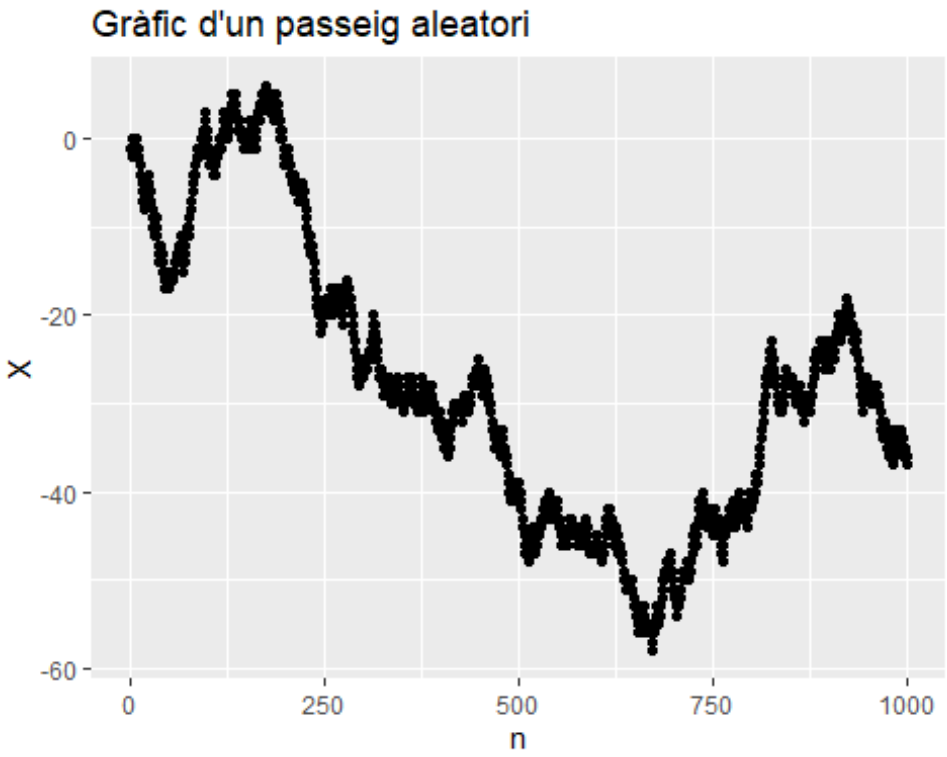
Simuleu una mostra d'aquest procés

Per a simular la mostra utilitzem la funció *sample()* que ens escollirà un número entre els valors $\{1, -1\}$ n vegades i guardarem aquestes variables aleatòries en un vector anomenat Y. A més a més, guardarem les sumes acomulatives de Y en un vector X.

```
n=1000
p=0.5

Y=sample(c(1,-1),n,prob=c(p,1-p),replace=TRUE)
X=cumsum(Y)
```

Per fer el gràfic dels valors obtinguts creem una seqüència que vagui del 1 al 1000 on seràn els nostres valors de n per a l'eix de les abscisses i utilitzem el vector X on tenim guardades les sumes acumulatives.

```
library(ggplot2)

n=seq(1:1000)
dt=data.frame(n,X)
ggplot(dt, aes(x = n, y=X)) +
  geom_point() + ggtitle("Gràfic d'un passeig aleatori")
```



El gràfic mostra un passeig aleatori per n=1000 on per a cada n la X augmenta o disminueix 1.

Apartat c

Descriviu el mètode de Montecarlo És un mètode per aproximar expressions matemàtiques. La principal idea del mètode de Montecarlo és la realització de moltes simulacions aleatòries. Per la llei dels grans nombres, la mitjana mostral convergeix a l'esperança poblacional.

Passos generals a seguir

1. Definir un domini de possibles inputs
2. Generem inputs aleatoris d'una distribució de probabilitat dins del domini
3. Calculem una mesura estadística dels inputs
4. Agregem els resultats

Com l'aplicaríem per calcular probabilitats d'un passeig aleatori

Com el passeig aleatori segueix una distribució binomial amb valors $\{1, -1\}$ i amb paràmetres n i p, podem simular aquest procès. El nostre domini de possibles inputs serà -1 i 1. Per generar els inputs aleatoris utilitzem la funció *sample()* que escull n vegades, amb probabilitat p, un número entre -1 i 1. A partir d'aquest vector, podem calcular les seves sumes cumulatives, que serà la nostra variable X_n, i a partir d'aquest nou vector podem estimar freqüències. Per la llei dels grans nombres, aquesta freqüència s'aproxima a una probabilitat.

Apartat d

Calculeu mitjançant Montecarlo quina és la probabilitat: $P([X_{1000} < 75])$

Per calcular la probabilitat farem us de la funció *sample()* per realitzar 1000 variables aleatòries amb valors $\{1, -1\}$. Una vegada hem creat les 1000 variables, si la suma d'aquestes és més petita que 75 el comptador augmentarà en 1.

Finalment, mostrem la nostra probabilitat fent el quocient entre el comptador i les n simulacions.

```
nsim=100
p=0.55
n=1000
P=0

for (j in 1:nsim){
  Y=sample(c(1,-1),n,prob=c(p,1-p),replace=TRUE)
  if (sum(Y)<75) P=P+1
}

P/nsim

## [1] 0.25
```

$P([X_{1000} < 75]) \approx 0.20$

Apartat e

Calculeu la següent probabilitat: $P([X_{12} < 7 \cap X_{534} < 127 \cap X_{1000} < 363])$

Tornem a realitzar els mateixos bucles que l'apartat anterior però en aquest cas la suma del vector Y ha de complir 3 condicions alhora i per tant en el if fem que valori les 3 condicions amb & (and).

```
nsim=100
p=2/pi
n=1000
Y=numeric(n)
P=0

for (j in 1:nsim){
  Y=sample(c(1,-1),n,prob=c(p,1-p),replace=TRUE)
  if (sum(Y[1:12])<7 & sum(Y[1:534])<127 & sum(Y)<365) P=P+1
}

P/nsim

## [1] 0.19
```

$P([X_{12} < 7 \cap X_{534} < 127 \cap X_{1000} < 363]) \approx 0.20$

Exercici 2

Apartat a

Demostreu que E(X) = Var(X) = λ

Per a l'esperança:

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x-1)!} e^{-\lambda} = \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda}$$

Fem un canvi de variable $z = x - 1$

$\lambda \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} e^{-\lambda} = \lambda \sum_{z=0}^{\infty} P(z) = \lambda$. Sabem que la suma de $P(z) = 1$.

Per a la variància:

$Var(X) = E(X^2) - [E(X)]^2$

$E(X^2) = E[X(X-1)] + E(X)$

$$E[X \cdot (X-1)] = \sum_{x=0}^{\infty} x\,(x-1) \cdot \frac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x-2)!} e^{-\lambda} = \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} e^{-\lambda}$$

Fem un canvi de variable $z = x - 2$

$\lambda^2 \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} e^{-\lambda} = \lambda \sum_{z=0}^{\infty} P(z) = \lambda^2$. Sabem que la suma de $P(z) = 1$.

Per tant, $E(X^2) = \lambda^2 + \lambda$

$Var(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

Apartat b

Calculeu l'estimació puntual i l'interval de confiança del Coeficient de Dispersió mitjançant el mètode de boostrap paramètric

Creem un vector x amb les

```
n_micro = c(0,1,2,3,4,5)
n_mostres = c(122,40,14,16,6,2)

x = c(rep(0,122),rep(1,40),rep(2,14),rep(3,16),rep(4,6),rep(5,2))
```

Tal i com hem demostrat al apartat a), l'esperança de la variable X és igual a la lambda. Per calcular lambda, sabem que és la mitjana de les dades.

```
E_x=mean(x)
```

En el mètode de boostrap paramètric tenim en compte la distribució de les dades, per tant utilitzem la funció *rpois()* per simular les n mostres.

```
nsim = 1000
CD.boot=numeric(nsim)
n=length(x)

for (i in 1:nsim){
  pr = rpois(n,E_x)
  CD.boot[i] = var(pr)/mean(pr)
}

quantile(CD.boot, probs=c(0.5,0.025,0.975))

##       50%      2.5%     97.5% 
## 0.9912119 0.8009172 1.1971008
```

Veiem que l'interval de confiança és (0.81,1.2) el qual correspont al CD, ja que en el cas d'una variable Poisson és $CD = \lambda/\lambda$, ja que $E(X) = Var(X) = \lambda$

Apartat c

Utilitzeu el mètode bootsrap no paramètric per estimar el CD

En aquest cas no podem estimar utilitzant cap distribució, ja que el mètode bootstrap no paramètric no assumeix cap distribució. En aquest cas, farem us de la variable *sample()* del vector x amb reemplaçament.

```r
CD.noboot = numeric(nsim)

for (i in 1:nsim){
  xB = sample(x,n,replace=T)
  CD.noboot[i] = var(xB)/mean(xB)
}

quantile(CD.noboot, probs=c(0.5,0.025,0.975))
```

```
##      50%     2.5%    97.5%
## 1.817913 1.545708 2.111296
```

Que podem dir si ho comparem amb els resultats anteriors

```r
var(x)/mean(x)
```

```
## [1] 1.832496
```

Quan ho comparem amb el mètode bootstrap paramètric veiem que els intervals de confiança són bastant diferent. En el primer cas, com coneixem la distribució del paràmetre fem l'estimació a partir d'aquesta i per tant el interval que ens dona s'apropa més al valor que ens hauria de donar (aproximadament 1). Quan fem el boostrap no paramètric desconeixem la distribució del paràmetre i al fer els intervals de confiança s'assemblen més al coeficient que obtenim amb les dades.

Apartat d

Calculeu l'estimació del mateix coeficient, però ara mitjançant el mètode de Jackknife

Tornem a no assumir cap distribució, però en aquest cas utilitzem el mètode Jackknife (eliminem la variable j-èssima en cada cas).

```r
cd=var(x)/mean(x); cd
```

```
## [1] 1.832496
```

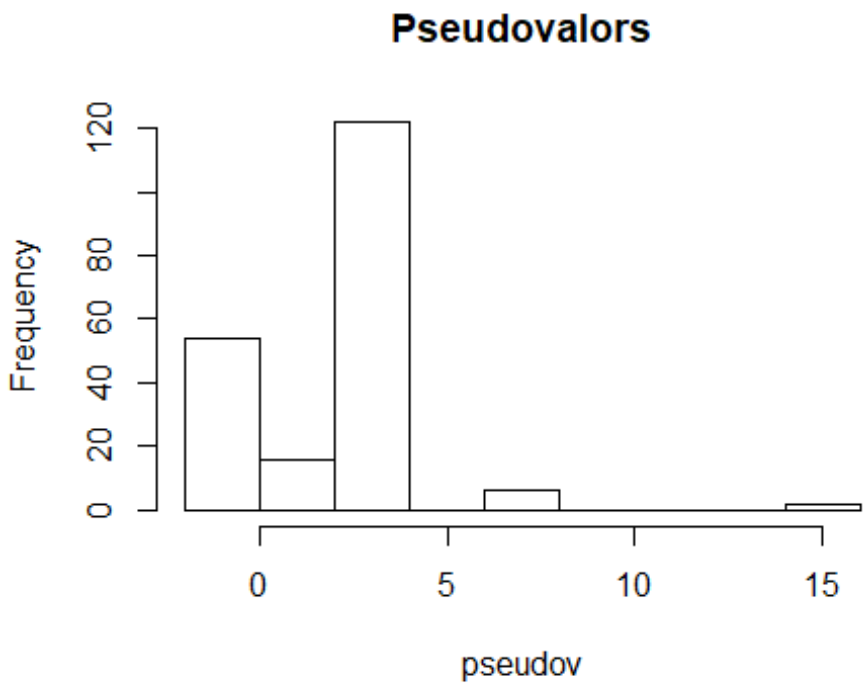
```r
n=length(x)
theta_jac = numeric(n)
pseudov = numeric(n)

for (i in 1:n){
  theta_jac[i] = var(x[-i])/mean(x[-i])
  pseudov[i] = n*cd-(n-1)*theta_jac[i]
}

est_jac <- mean(pseudov)
var_jac <- sd(pseudov)/sqrt(n)
c(est_jac, est_jac - 1.96*var_jac, est_jac + 1.96*var_jac)
```

```
## [1] 1.833644 1.548521 2.118767
```

Com no sabem la distribució, el seu interval de confiança també s'aproxima al valor obtingut amb les dades. Aquest interval és molt semblant al obtingut amb el cas no paramètric.

Valoració dels pseudovalors

```
hist(pseudov, main="Pseudovalors")
```




```
table(pseudov)

## pseudov
## -0.982614752715278 -0.539885721060045    1.3034639976492
2.57707493697444
##                 14                  40                  16
122
##   6.37442192281458   14.2878772126043
##                  6                   2
```

Com tenim molts casos de la mostra repetits, ja que es tracta d'un recompte, amb els pseudovalors succeeix el mateix. Els tenim repetits on hi ha dos casos molt alts comparats amb els altres.

Apartat e

Tenint en compte el mètode Jackknife, podem assumir que les dades provenen d'una distibució de Poisson?

No podem assumir que són Poisson, ja que el CD és bastant diferent al valor teòric per una Poisson (aproximadament 1).

Us sembla correcte aplicar el mètode boostrap paramètric

No, ja que amb el mètode Jack acabem de demostrar que no podem assumir que són Poisson i per tant no sería correcte aplicar un mètode basat en la distribució del paràmetre que desconeixem.