Løsningsforslag øving 11, ST1301

Leveringsfrist: Mandag 14. april, kl 12:15 på e-post til Yngvild. Teller 20% på karakteren. Øvingen besvares individuelt og leveres med studentnummer og uten navn.

Oppgave 1 Anta at et antall arter $i = 1, 2, \ldots, s$ utgjør andeler $p_1, p_2, \ldots, p_s$ av et dyresamfunn. Et mye brukt mål på et slikt samfunns artsmangfold (diversitets) er Simpsons indeks

$$H_S = 1 - \sum_{i=1}^{n} p_i^2. \tag{1}$$

Vis at denne indeksen er lik sannsynligheten for at to tilfeldig valgte individ trukket fra samfunnet ikke tilhører samme art.

Sannsynligheten for å trekke art i to ganger blir p_i^2. Total sannsynligheten for å trekke samme art blir dermed $\sum_{i=1}^{s} p_i^2$ og sannsynligheten for å ikke trekke samme art (komplementet) blir da $1 - \sum_{i=1}^{n} p_i^2 = H_S$. Dette betyr blant annet at H_S øker med antall arter i samfunnet og avtar hvis enkelte arter dominerer (utgjør en stor andeler).

Dersom vi trekker et utvalg invidid av størrelse N fra samfunnet følger det at antallet av art $1, 2, \ldots, s$, vektoren $(X_1, X_2, \ldots, X_s)$, er multinomisk fordelt.

Anta at det i et samfunn er $s = 10$ arter tilstede og at artene utgjør andelene $0.5, 0.2, 0.1, 0.05, 0.05, 0.05, 0.02, 0.01, 0.01, 0.01$. Hva er da den sanne verdien av H_S?

Parameteren H_S blir en funksjon av parameterne $p_1, \ldots, p_s$. Det kan være hensiktsmessig å implementere dette som en funksjon vi kan bruke senere.

```
## Kontrakt: vektor -> flyttall
## Hensikt: Beregne Simpsons diversitetsindeks gitt samfunnets
## strukter p
##
HS <- function(p) {
  1 - sum(p^2)
}
```

Vi kan så beregne H_S for p_i'ene gitt i oppgaven.

```
> p.oppg <- c(0.5,0.2,0.1,0.05,0.05,0.05,0.02,0.01,0.01,0.01)
> HS(p.oppg)
[1] 0.6918
```

Beregn forventningsfeilen til estimatoren

$$\hat{H}_S = 1 - \sum_{i=1}^{s} \left(\frac{X_i}{N} \right)^2 \tag{2}$$

for utvalgstørrelser $N = 10$, $N = 100$ og $N = 1000$. Du vil få bruk for funksjonen `rmultinom` i R for å simulere fra multinomisk fordeling.

Det kan her være hensiktsmessig først å implementere $\hat{H}_S$ som egen egen funksjon.

```
## Kontrakt : vektor -> flyttall
##
## Hensikt: Beregne estimat av H_S gitt antall av arter av ulike arter
## i utvalget representert ved vektoren X
##
HShat <- function(X) {
   N <- sum(X)
   1 - sum((X/N)^2)
}
```

Funksjon for å beregne forventningsfeilen $E(\hat{H}_S) - H_S$ kan så utformes på følgende måte. Legg merke til bruken av argumentet `est`.

```
## Kontrakt: funksjon, vektor, heltall -> flyttall
##
## Hensikt: Beregne forventingsfeil til estimator gitt ved funskjon est
## gitt samfunnets strukter p og utvalgsstørrelse p
##
bias <- function(est,p,N) {
  sannHS <- HS(p)
  HSsim <- rep(NA,1000)
  for (i in 1:1000) {
     X <- rmultinom(n=1,size=N,prob=p)
     HSsim[i] <- est(X)
  }
  mean(HSsim)-sannHS
}
```

Forventningsfeilen kan nå beregnes på følgende måte.

```
> bias(HShat,p.oppg,10)
[1] -0.06988
> bias(HShat,p.oppg,100)
[1] -0.0069526
> bias(HShat,p.oppg,1000)
[1] -0.00022953
```

Ikke uventet avtar forventningsfeilen med utvalgstørrelsen N.

Undersøk også forvetningsfeilen til estimatoren

$$\hat{H}'_S = 1 - \sum_{i=1}^{s} \frac{X_i(X_i - 1)}{N(N-1)} \tag{3}$$

for samme verdier av N.

Fordelen ved å la funskjonen som beregner estimater være innargument til funskjonen `bias` er at vi nå kun trenger å implementere $\hat{H}'_S$ som en funksjon i R for å beregne dennes forventnignsfeil uten å måtte skrive om `bias`.

```
HShatmerket <- function(X) {
   N <- sum(X)
   1 - sum((X/N)*(X-1)/(N-1))
}
> bias(HShatmerket,p.oppg,100)
[1] 2.222222e-05
> bias(HShatmerket,p.oppg,1000)
[1] -0.0006266667
> bias(HShatmerket,p.oppg,1000)
[1] 5.661662e-06
```

Hvilken estimator ser ut til å være å foretrekke?

Vi ser at $\hat{H}'_S$ ser ut til å ha minst forventingsfeil for de parameterverdiene vi har brukt og dette indikerer at denne estimatoren er å foretrekke.

Kan vi estimere H_S uten å kjenne det totale antallet arter s i samfunnet?

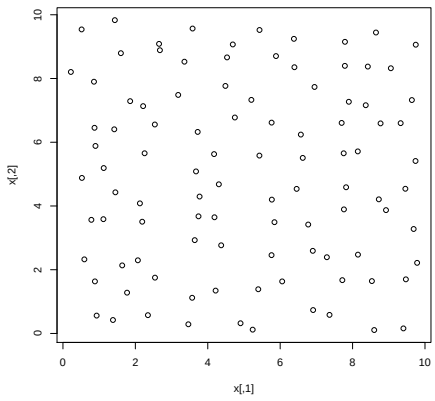
De artene vi ikke observerer, d.v.s. de som opptrer i antall $X_i = 0$ i utvalget, gir ikke noe bidrag til summene i (2) og (3) og vi kan dermed estimerer H_S uten å vite noe om det totale antall arter s. Det kan bemerkes at dataene i praksis altså ikke vil være multinomisk fordelt fordi vi har en form for sensurering ved at kun arter i med $X_i \geq 1$ observeres.

Oppgave 2 Anta at vi observerer 100 trær i et plantefelt på 10 ganger 10 meter. Les inn koordinatene til trærne med kommandoen

```
x<-matrix(scan(file="http://www.math.ntnu.no/~jarlet/skog.dat"),ncol=2)
```

Objektet `x` blir en 100×2 matrise som inneholder koordinatene til hvert tre i skogen. Lag et plot av skogen med kommandoen `plot(x)`.

Plottet blir seende slik ut.



En mulig modell er å tenke seg plantefeltet beskrevet som en Poisson-prossess i planet med rate λ. Hva forutsetter denne modellen?

En Poisson-prosess i planet er vanligvis homogen, d.v.s. ingen romlig variasjon i intensiteten λ, to eller flere hendelser kan ikke inntreffe i samme punkt, og hendelser inntreffer uavhengig av hverandre.

Er antakelsene biologisk realistiske?

Det kan tenkes at λ varierer romlig p.g.a. av variasjon i vekstforholdene (d.v.s. at prosessen er inhomogen). Vi kan også ha avhengighet ved at nærtliggende individ etter hvert (men kanskje ikke umiddelbart etter spiring) vil konkurrere om ressurser som næring, vann og lys. I tillegg kan ikke

ulike inidivid befinne seg svært nær hverandre slik som i Poisson-prossess fordi ulike individ har en viss fysisk utstrekning. Dette innebærer en annen for for avhengighet. Antakelsen om at hendelser ikke kan intreffe i samme punkt er imidlertid oppfyllt.

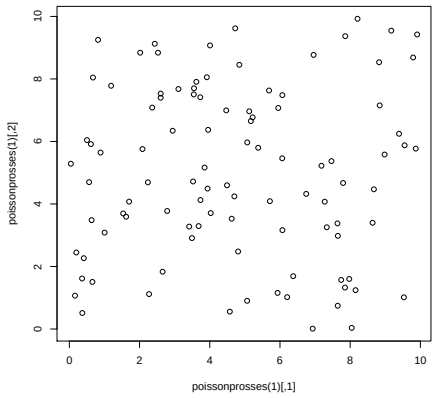
Dersom plantefeltet er beskrevet ved en Poisson-prossess er antall invidid i kvadratet Poissonfordelt med parameter λA der A er arealet av kvadratet. Hva blir estimatet av λ når plantefeltet er på 10 ganger 10 meter?

Estimatet blir antall hendelser $X = 100$ delt på arealet $A = 100m^2$, d.v.s. $\hat{\lambda} = 100/100m^2 = 1m^{-2}$.

Lag en funksjon som simulerer en ny realisasjon av plantefeltet under Poisson-prossess-antakelsen, for en gitt verdi av λ og som returnerer koordinatene i form av en $(N \times 2)$-matrise der N er antall trær. Hint: Simulerer først antall hendelser (hendelser) N i plantefeltet. Gitt dette antallet er da posisjonene uniformt fordelte på kvadratet på 10 ganger 10 meter.

```
## Hensikt: Simulerer en romlig poissonprossess i et kvadrat på 10x10 meter
## med intensitet lambda og returnere koordinatene til hendelsene i form
## av en N x 2 matrise.
##
poissonprosess <- function(lambda=1) {
  X <- lambda*10*10
  koordinater <- runif(2*X,0,10) # uniformt fordelte x- og y-koordinater
  matrix(koordinater,ncol=2) # ordne disse i en matrise.
}
plot(poissonprosess(1))
```

Lag et plot av en realisasjon og sammenlign visuelt med den observerte skogen. Ser du noen forskjell mellom det simulerte og observerte mønsteret?



Vi ser at punktene i den observerte skogen er jevnere fordelt mens det i det simulerte skogen er en mer klumpvis fordeling.

Vi ønsker å bruke Poisson-prossess-antakelsen som en null-hypotese og teste denne. Som testobservator T velger vi å bruke minste avstand mellom to nabotrær. Dette blir en funksjon av koordinatene til alle trærne som lett kan beregnes ved hjelp av funksjonene `dist` og `min` i R. Hvilken verdi T^* tar testobservatoren for den observerte plantefeltet?

Igjen er det hensiktsmessig å implementer T (testobservatoren) som en funksjon i R.

```
## Hensikt: Beregne minste avstand mellom to hendelser i en
## realisasjon av en romlig punktprosess.
##
Tfunk <- function(x) {
  min(dist(x))
}
> Tfunk(x)
[1] 0.2004773
```

Ingen trær er m.a.o. nærmere hverandre enn 20,04 cm.

Simulere mange realiasjoner av plantefeltet under H_0, beregn tilhørende verdier av testobservatoren T, og beregn til slutt testens p-verdi, $P(T > T^*)$. Kommenter resultatet.

```
pverdi <- function(xobs) {
  Tobs <- Tfunk(xobs)
  teller <- 0
  for (i in 1:1000) {
    x <- poissonprosess(1)
    if (Tfunk(x)>Tobs)
       teller <- teller + 1
  }
  teller/1000
}
> pverdi(x)
[1] 0.003
```

Vi ser at sannsynligheten for den observerte eller større verdier av T er svært liten (0.003) og vi kan dermed forkaste hypotesen H_0 om at skogen kan beskrives som en Poissonprosess. Det er ikke gitt at testobservatoren vi har brukt er noen god testobservator. Skulle vi sett på styrken vi får ved bruk av ulike alternative testobservatorer måtte vi simulert også under alternativ hypotese H_1.

Det nevnes at dataene i oppgaven var simulert ved å plassere trærne i et 10 x 10 rutenett tillagt uniformt fordelte avvik på følgende måte

```
x <- matrix(c(rep(0:9,10)+runif(100,0.05,.95),
              rep(0:9,rep(10,10))+runif(100,0.05,0.95)),
            ncol=2)
```

som studentnr 698959 helt riktig var inne på.