Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsforslag

Bioberegninger, ST1301

Mandag 21. mai, 2007

Oppgave 1

a)
```
> x <- c(180,185,190,165,160,155)
> y <- c(1,1,2,2,2,3)
> x[x<mean(x)]
[1] 165 160 155
> mean(x[y==2])
[1] 171.6667
> length(y==2)
[1] 6
> length(x[y==2])
[1] 3
```

Merk at den logiske operatoren `==` virker elementvis på vektoren `y` slik at uttrykket `y==2` blir en (logisk) vektor med samme lengde som `y`, altså 6.

b)
```
> funk <- function(x) {
+   n <- length(x)
+   teller <- 0
+   for (i in 2:(n-1)) {
+      if (x[i]==x[i+1] & x[i]==x[i-1]) {
+         teller <- teller + 1
+      }
+   }
+   teller
+ }
> funk(c(0.5,0.5,1,1,1,2,2,1,2,0,0,0,1,1,0,0))
[1] 2
> funk(c(0.5,0.5,1,1,1,2,2,1,2,0,0,0,0,1,1,0,0))
[1] 3
```

Oppgave 2

a) Null-hypotesen vi har testet er at korrelasjonen mellom variablene alder og vekt i populasjonen er null. Siden p-verdien over er større enn $\alpha = 0.05$ kan vi ikke forkaste null-hypotesen. Det betyr at null-hypotesen kan være riktig men manglende forkastning kan også skyldes lav teststyrke (type II feil).

b) Korrelasjons-testen forutsetter normalfordelte data. Her ser alder ut til å ha en skjev fordeling slik at normalfordelingsantakelsen ikke er oppfyllt.

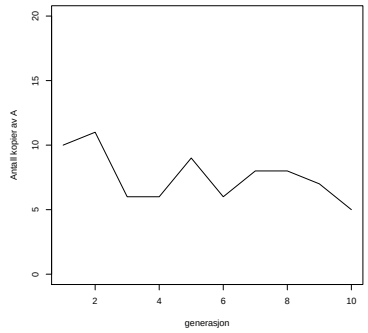
c) En mye brukt fordeling for å modellere levetider er eksponensialfordelingen. Den ser også ut til å passe godt med de observerte dataene i dette tilfelle.

d)
```
## prob.type.I.feil : flyttall, flyttall, flyttall, flyttall, heltall, heltall
##                       -> flyttall
##
## Hensikt: Estimere sannsynligheten for type I feil i korrelasjonstest
## når den ene variabelen er eksponensiellt (og ikke normal-) fordelt, gitt
## utvalgsstørrelse n, parameterverdier, og signifikansnivå alpha.
##
prob.type.I.feil <- function(lambda,mu,sigma,alpha,n,nsim=1000) {
  teller <- 0
  t.kritisk <- qt(df=n-2,p=alpha/2,lower.tail=FALSE)
  for (i in 1:nsim) {
    x <- rnorm(n,mu,sigma)
    y <- rexp(n,lambda)
    t.sim <- cor.test(x,y)$statistic
    if ((t.sim>t.kritisk)|(t.sim< -t.kritisk))
       teller <- teller + 1
  }
  teller/nsim
}
```

Oppgave 3

a)
```
drift.sim <- function(X1,N,n) {
  X <- rep(NA,n)
  X[1] <- X1
  for (t in 2:n)
     X[t] <- rbinom(n=1,prob=X[t-1]/(2*N),size=2*N)
```

```
  X
}
```



b) Et heterozygot individ kan dannes på to måter, enten ved at genvarianten A arves fra mor eller fra far. Hver av disse hendelsene inntreffer med sannsynlighet $p_t(1-p_t)$ slik at den totale sannsynligheten for at et individ er heterozygot (og frekvensen av heterozygoter i populasjonen) blir $2p_t(1-p_t)$.

c)

```
forventet.heterozygotfrekvens <- function(N,X1,n,nsim=1000) {
  H <- rep(NA,nsim)
  for (i in 1:nsim) {
    p <- (drift.sim(X1,N,n)[n])/(2*N) # frekvensen i siste generasjon
    H[i] <- 2*p*(1-p) # tilsvarende realisasjon av heterozygotfrekvensen
  }
  mean(H)
}
```

d) Vi har at heterozygotfrekvensen, funksjonen $H_t = 2p_t(1-p_t)$ av genfrekvensen p_t, har sitt maksimum $H_t = 1/2$ i $p_t = 1/2$. Så snart vi får genetisk drift vekk fra $p_t = 1/2$ må altså H_t avta. Dermed må også $E(H_t) < H_1$.

Et litt mer stringent bevis: Vi har at $H_t < 1/2$ for alle $p_t \neq 1/2$. Fra definisjon av

forventningsverdi får vi at

$$\begin{aligned} EH_t &= \sum_h hP(H_t = h) \\ &= \sum_{h \neq 1/2} hP(H_t = h) + 1/2P(H_t = 1/2) \\ &< \sum_{h \neq 1/2} 1/2P(H_t = h) + 1/2P(H_t = 1/2) \\ &= 1/2 = H_1, \end{aligned} \tag{1}$$

når $p_1 = 1/2$.

e) Genfrekvensen etter n generasjoner er en stokastisk variabel med en viss forventning (oppgitt å være lik p_1 i oppgaven) og varians. Heterozygotfrekvensen er en funksjon av denne. Vanlige regneregler fra sannsynlighetsteorien gir at

$$\begin{aligned} EH_t &= E2p_t(1 - p_t) \\ &= 2Ep_t - 2E(p_t^2) \\ &= 2Ep_t - 2[\operatorname{Var} p_t + (Ep_t)^2] \\ &= 2p_1 - 2[\operatorname{Var} p_t + (p_1)^2] \\ &= 2p_1(1 - p_1) - 2\operatorname{Var} p_t \\ &= H_1 - 2\operatorname{Var} p_t \\ &< H_1 \end{aligned} \tag{2}$$

siden $\operatorname{Var} p_t > 0$ (p_t er en stokastisk variabel med positiv varians). Altså vil forventet heterozygotfrekvens alltid avta med tiden t.