Øving 12, ST1301

Oppgave 1 En to-utvalgs t-test forutsetter at observasjonene i hvert utvalg $X_1, X_2, \ldots, X_n$ og $Y_1, Y_2, \ldots, Y_m$ er uavhengige normalfordelte variable. Hvis testen oppfører seg slik den skal også dersom disse forutsetningene ikke er oppfyllt sier vi at testen er robust. Dersom den nominelle (ønskede) sannsynligheten for type I feil (sannsynligheten for å forkaste H_0 gitt at H_0 er sann, signifikansnivået) er valgt lik α, innebærer dette at vi ønsker at den reelle sannsynligheten for type I feil skal ligge nære α også dersom dataene i virkeligheten ikke er normalfordelte.

a) Anta at vi observerer følgende to utvalg:

A: 0.41 0.61 0.85 2.19 0.92 0.26 0.38 1.06 0.43 0.44

B: 5.60 3.28 3.67 0.75 0.57 2.32 6.40 1.60 1.13 0.86

Anta at dataene er normalfordelte samme varians i hvert utvalg og med forventninger μ_A og μ_B. Undersøk om $\mu_A < \mu_B$ ved hjelp av funksjonen `t.test` i R. Husk å spesifisere antakelsen om like varianser i hver gruppe.

b) Programmer en funksjon som beregner den reelle sannsynligheten for type I feil dersom dataene i virkeligheten er eksponentielt fordelte ved å simulere 1000 realisasjoner av de to utvalgene. Tilhørende simulerte verdier av testobservatoren kan beregnes ved å referere til `$statistic`-komponenten i listen som returneres av `t.test`.

Bruk funksjonen til å beregne den reelle sannsynligheten for type I feil dersom størrelsene på de to utvalgene er 2, 5, 10, og 20.

c) Ser dataene i punkt a) normalfordelte ut? Vil du vurdere test-resultatet i punkt a) som pålitelige ut i fra simuleringsresultatene i punkt b)?

Oppgave 2 *Leslie-matriser og alderstrukturerte modeller*

Anta at vi har en populasjon hvor reproduksjon inntreffer i en kort "puls" på et gitt tidspunkt hvert år, f.eks. om våren. Anta videre at høyeste alder er 4 år, at overlevelsen fra alder $i-1$ til alder i er gitt ved s_i, og at $s_1 = 0.2$, $s_2 = 0.5$, $s_3 = 0.7$ og $s_4 = 0.5$. Vi ser kun på hunnene i populasjonen. (Dette er vanlig å gjøre for organismer hvor hannene ikke bidrar til oppfostring av avkom.) Anta at hunner ved alder i i snitt får f_i hunnlige avkom og at $f_1 = 2$, $f_2 = 4$, $f_3 = 4$ og $f_4 = 2$.

Hvis vi lar søylevektoren $\mathbf{N}_t$ representere antall individer av ulik aldere ved tidspunkt t kan vi nå konstruere en Lesliematrise $\mathbf{A}$ slik at

$$\mathbf{N_{t+1}} = \mathbf{A}\mathbf{N_t} \tag{1}$$

på to måter, enten ved å la elementene i $\mathbf{N_t}$ representere antall individer i de ulike aldersklassen rett *forut* for reproduksjon (såkalt prebreeding census) eller rett etter reproduksjon (postbreeding census).

For prebreeding census vil elementene i $\mathbf{N}_t$, antallet individer $N_{1,t}, N_{2,t}, \ldots, N_{4,t}$ i *aldersklasse* 1, 2, 3 og 4, være individer som er henholdsvis nesten ett, to, tre og fire år gamle ved tidspunkt t. Census (opptelling av populasjonen) er i dette tilfelle etterfulgt av reproduksjon og så overlevelse før neste census. Vi har da at

$$N_{i+1,t+1} = s_{i+1} N_{i,t}. \tag{2}$$

Antall individer i første aldersklasse ved tidpunkt $t+1$ blir

$$N_{1,t+1} = \sum_{i=1}^{4} s_1 f_i N_{i,t}, \tag{3}$$

fordi nyfødte også må overleve i nesten ett år fram til neste census. Skriver vi ligning (2) og (4) på matriseform som i (1) får vi at

$$\mathbf{A} = \begin{bmatrix} s_1 f_1 & s_1 f_2 & s_1 f_3 & s_1 f_4 \\ s_2 & 0 & 0 & 0 \\ 0 & s_3 & 0 & 0 \\ 0 & 0 & s_4 & 0 \end{bmatrix}. \tag{4}$$

For postbreeding census vil elementene i $\mathbf{N}_t$, antallet individer $N_{1,t}, N_{2,t}, \ldots, N_{4,t}$ i *aldersklasse* 1, 2, 3 og 4, være individer som er henholdsvis såvidt null (nyfødte), ett, to og tre år gamle ved tidspunkt t. Nå er census etterfulgt av overlevelse og så reproduksjon rett forut for neste census. Nå vil antall individer i andre aldersklasse ved tid $t+1$, $N_{2,t+1}$, altså antall individer som såvidt er ett år gamle ved tid $t+1$, være antall i første aldersklasse (nyfødte) ved tid t, $N_{1,t}$ multiplisert med overlevelsen fra alder 0 til 1, s_1. Generellt blir

$$N_{i+1,t+1} = s_i N_{i,t}. \tag{5}$$

Merk forskjellen fra (2). Videre vil f.eks. andre aldersklasse ved tid t, $N_{2,t}$, d.v.s. individer som er såvidt ett år gamle bidra med $s_2 f_2$ nyfødte hver fordi de først må overleve til alder 2 hvor de får f_2 avkom hver i snitt. Generellt blir dermed

$$N_{1,t+1} = \sum_{i=1}^{4} s_i f_i N_{i,t}. \tag{6}$$

Legg igjen merke til forskjellen fra (3).

a) Hvordan blir matrisen $\mathbf{A}$ nå seende ut uttrykt ved $s_1, \ldots, s_4$ og $f_1, \ldots, f_4$?

b) Programmer to funksjoner som beregner $\mathbf{A}$ for henholdsvis prebreeding og postbreeding census. La to vektorer som inneholder overlevelsesparameterne $s_1, \ldots, s_4$ og fekunditetene $f_1, \ldots, f_4$ være innargument til funksjonene. Hint: Funksjonene `matrix`, `diag`, `bind`, `rbind` vil kunne være nyttige.

c) Beregn vekstraten til populasjonen basert på hver av Leslie-matrisene ved å se på første egenverdi (beregnes v.h.a. funksjonen `eigen` i R.) Kommenter resultatet.

d) Beregn også de de stabilie aldersfordelingene ved å se på første egenvektor. Hvorfor blir disse forskjellige for post- og prebreeding census?

e) Ved et gitt tidspunkt t vil antall i hver aldersklasse være gitt ved $\mathbf{N}_t = \mathbf{A}^t\mathbf{N}_0$. Anta at populasjonen bare består 100 individer i alderklasse 1 ved tidspunkt $t = 0$. Beregn først antall i hver alderklasse ved tidspunkt $t = 6$ ved hjelp av matrisemultiplikasjon.

f) I stedet for å beregne $\mathbf{A}^t$ ved gjentatt matrisemultiplikasjon slik som over kan dette også gjøres ved å diagonalisere $\mathbf{A}$. Det kan vises at vi kan skrive matrisen $\mathbf{A}$ på formen $\mathbf{A} = \mathbf{PDP}^{-1}$ hvor $\mathbf{P}$ er en $n \times n$ matrise satt sammen av egenvektorene til $\mathbf{A}$ og $\mathbf{D}$ er en $n \times n$ diagonal matrise med elementer lik egenverdiene til $\mathbf{A}$. Denne framgangsmåten forutsetter at $\mathbf{A}$ er diagonaliserbar.

Bruk diagonaliseringen til å beregne antall individer i hver alderklasse ved $t = 6$ og $t = 50$. Hint: Husk at $\mathbf{A}^t = (\mathbf{PDP}^{-1})^t = \mathbf{PD}^t\mathbf{P}^{-1}$ og at $\mathbf{D}^t$ selv er diagonal.

Oppgave 3 *Simulering av stokastiske prosesser. Genetisk drift.*

Vi betrakter populasjon av en diploid organisme (kromosomene opptrer i par) med konstant populasjonsstørrelse N. I følge den idealiserte Wright-Fisher modellen er antall kopier av et allel (en variant av et gen) i generasjon $t+1$, X_{t+1}, gitt at X_t kopier av allelet er tilstede i generasjon t ($0 \leq X_t \leq 2N$) binomisk fordelt med parametere $p_t = X_t/(2N)$ og $n = 2N$. Vi har alstå at

$$X_{t+1} \sim \text{bin}(\frac{X_t}{2N}, 2N) \tag{7}$$

Dermed er $E(X_{t+1}|X_t) = X_t$ og $\text{Var}(X_{t+1}|X_t) = 2Np_t(1-p_t)$. Ser vi på allelfrekvensen $p_t = X_t/(2N)$ har vi at $E(p_{t+1}|p_t) = p_t$ og

$$\text{Var}(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{2N} \tag{8}$$

I forventingsverdi vil altså frekvensen av et allelet være den samme som i forrige generasjon men frekvensen vil endre seg; vi har såkalt genetisk drift. Vi ser av (8) at jo mindre populasjonsstørrelsen er, jo mer vil allelfrekvensen endre seg som følge av genetisk drift.

Programmer en funksjon `drift` som simulerer en realisasjon av prossesen over, d.v.s. antall kopier av allelet $X_1, X_2, \ldots, X_m$ i generasjon $1, 2, \ldots, m$. La populasjonsstørrelsen N, antall kopier av allelet i generasjon 1, X_1, og antall generasjoner m være innargumenter.

Lag et plot av 10 realisasjoner av prosessen ved hjelp av gjentatte kall til `drift`, `plot` og `lines`. La $X_1 = 5$, $N = 10$ og $m = 50$.

Programmer en funksjon som simulerer 1000 realisasjoner av prossesen over (ved hjelp av gjentatte kall til `drift`) og som estimerer sannsynligheten for at $p_m = 1$ gitt at $X_1 = 1$, $N = 10$ og $m = 200$. Hva blir sannsynligheten i forhold til allelefrekvensen p_t i første generasjon $t = 1$. Virker resultatet rimelig?

Oppgave 4 *Stokastiske prosesser i økologi*

La N_t være antall individer i en gitt populasjon ved tidspunkt t. Anta at N_{t+1} gitt N_t er normalfordelt med forventning

$$E(N_{t+1}|N_t) = N_t + rN_t(1 - \frac{N_t}{K}), \tag{9}$$

og varians

$$\mathrm{Var}(N_{n+t}|N_t) = \sigma_e^2 N_t^2. \tag{10}$$

Dette er en stokastisk versjon Lotka-Volterra modellen med miljøstokastisitet hvor parameteren σ_e^2 kalles miljøvarians. Parameteren r er vekstraten til populasjonen og K bæreevnen.

Lag først en funksjon som simulerer en realisasjon av prosessen $\{N_t\}$ fra $t = 1$ til $t = 1000$ gitt $N_1 = K$ og returner en vektor som inneholder $N_1, N_2, \ldots, N_t$ som funksjonsverdi. La parameterne r, K, og σ_e^2 være argument.

Lag plot av noen realisasjoner av prosessen for $K = 100$, $r = 0.1$, og $\sigma_e^2 = 0.05$. På hvilken måte endrer populasjonssvingningene seg dersom vi endrer vekstraten r samtidig som andre parametere holdes fast? Tilsvarende, hva skjer om vi endrer på miljøvariansen σ_e^2? (Det kan være hensiktsmessig å velge f.eks. `ylim=c(0,300)` i alle plottene for lettere å kunne sammenligne resultatene visuelt.)

Oppgave 5 *Kaos*

Betrakt rekursjonsligningen

$$x_{t+1} = f(x_t) = cx_t(1 - x_t) \tag{11}$$

hvor $0 < c < 4$. Hvis x_t går mot en likevekt x^* når $t \to \infty$ må

$$x^* = cx^*(1 - x^*), \tag{12}$$

som medfører at

$$x^* = 1 - \frac{1}{c}. \tag{13}$$

For x_t i nærheten av x^* har vi tilnærmet at

$$x_{t+1} = f(x^*) + f'(x^*)(x_t - x^*) \tag{14}$$

Siden $f(x^*) = x^*$ kan vi skrive (14) på formen

$$x_{t+1} - x^* = f'(x^*)(x_t - x^*). \tag{15}$$

Vi ser at dette medfører at avviket fra x^*, $x_t - x^*$ vil konvergere mot null (likevektspunktet x^* er stabilt) hvis

$$-1 < f'(x^*) < 1 \tag{16}$$

Fra (11) og (13) får vi at

$$f'(x^*) = 2 - c, \tag{17}$$

som innsatt i (16) medfører at

$$1 < c < 3 \tag{18}$$

for at x^* skal være et stabilt likevektspunkt.

Vi skal nå undersøke hva som skjer for $3 < c \leq 4$. Programmer en funksjon som tar c, x_1, og m som innargumenter og som beregner $x_1, x_2, \ldots, x_m$ og returner denne tallfølgen som funksjonsverdi. Lag så plot av tallfølgen for $m = 50$, $x_1 = .51$, og forskjellige valg av c lik 3.1, 3.5, 3.7.

Se spesielt på det det tilfelle at $c = 3.7$ og plot tallfølgen for to nesten identiske startverdier, f.eks. 0.51 og 0.511. Vil det være mulig å forutsi posisjonen til x_m gitt at vi ikke kjenner x_1 med fullstendig nøyaktighet?

Vi har sett at vi enten får stabile grensecykler eller kaos for verdier av $c > 3$. For å få bedre oversikt over oppførselen til prosessen for alle verdier av c skal vi lage et såkalt bifurkasjonsgraf. Denne grafen skal vise hvilke tilstander x_t er innom når prosessen har nådd en grensecykel eller kaotisk oppførsel for ulike verdier av c. Dette kan i praksis gjøres ved å lage en for-løkke som løper igjennom en sekvens verdier for c (f.eks. 500 verdier fra 1 til 3). For hver c-verdi itereres prosessen 200 ganger og de siste 100 verdiene legges til som punkter i plottet med `points`. Bruk `pch="."` som tilleggsargument til `points`.