TMA4245 Statistikk
Vår 2017

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Innlevering 3

Dette er den første av to innleveringer i blokk 2. Denne øvingen skal oppsummere pensum forelest i uke 8-11. Øvingen handler om funksjoner av stokastiske variabler (kap. 7 i lærebok og notat om ordningsvariabler), estimering og konfidensintervall (kap. 8 og 9 i læreboka). Alle deloppgaver teller like mye.

Oppgave 1 Juletrelysene

Vi ser på en lenke med 36 juletrelys der lysene er seriekoblet. Det betyr at hvis ett lys slukner så slukner hele lyskjeden. Anta videre at levetiden til lys nr i, X_i, $i = 1, \ldots, 36$, er eksponentialfordelt med forventningsverdi μ timer, og at levetiden til de ulike lysene er uavhengig av hverandre. La U være en stokastisk variabel som angir levetiden til lyslenken.

a) Hva er sammenhengen mellom U og $X_1, X_2, \ldots, X_n$? Finn fordelingen til levetiden U til lyslenken. Finn også forventet levetid til lenken, $\mathrm{E}(U)$, og angi numerisk verdi når $\mu = 5000$ timer.

b) Når lyslenken fungerer bruker den 72 W, dvs. at i løpet av en time vil den bruke 0.072 kWh. Hvilken fordeling har energiforbruket Y til lyslenken over hele dens levetid? (Hint: $Y = 0.072U$). Finn forventet energiforbruk for lyslenken når $\mu = 5000$ timer.

Oppgave 2

I forkant av et stortingsvalg blir det gjennomført en meningsmåling der et representativt utvalg av velgerne blir spurt om de ønsker et regjeringsskifte eller ikke. Anta at andelen av velgerne som ønsker et skifte er p, og la X være antall personer blant n spurte som svarer JA på spørsmålet "Ønsker du et regjeringsskifte ved høstens valg?".

Anta at to aviser på en bestemt dag presenterer resultater fra to meningsmålinger, gjennomført av hvert sitt meningsmålingsinstitutt, Byrå A og Byrå B. La n_1 være antall spurte og X_1 antall som svarer JA i målingen fra Byrå A, og n_2 og X_2 tilsvarende størrelser for Byrå B. Vi antar at X_1 er binomisk fordelt med parametre n_1 og p, og X_2 er binomisk fordelt med parametre n_2 og p, og at X_1 og X_2 er uavhengige.

Vi ønsker å estimere p ved å kombinere resultatene fra de to målingene. To aktuelle estimatorer

er

$$\begin{aligned} \hat{P} &= \frac{1}{2}\left(\frac{X_1}{n_1} + \frac{X_2}{n_2}\right) \quad \text{og} \\ P^* &= \frac{X_1 + X_2}{n_1 + n_2}. \end{aligned}$$

a) Finn forventning og varians til hver av de to estimatorene $\hat{P}$ og P^*.

Dersom $n_1 = 500$ og $n_2 = 1000$, hvilken estimator vil du da velge? Begrunn svaret.

Anta nå at $n_1 = n_2 = n$, slik at X_1 og X_2 er uavhengige og binomisk fordelte, med *samme* parametre p og n. Dette medfører at

$$\hat{P} = P^* = \frac{X_1 + X_2}{2n}.$$

Utled et tilnærmet 95% konfidensintervall for p ved å bruke at fordelingen til

$$\frac{\hat{P} - p}{\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}}$$

er tilnærmet standard normalfordelt.

Et tredje meningsmålingsinstitutt, Byrå C, har annonsert at de snart kommer med resultater fra en tilsvarende måling med n_3 spurte. La X_3 være antall som svarer JA på spørsmålet om regjeringsskifte i målingen fra Byrå C, og anta at X_3 er uavhengig av X_1 og X_2. Vi vil nå bruke resultatene fra Byrå A og Byrå B til å predikere hvor mange som svarer JA i den nye målingen. Vi antar i resten av oppgaven at $n_1 = n_2 = n_3 = n = 1000$, og at observerte verdier for X_1 og X_2 er $x_1 = 645$ og $x_2 = 692$.

b) La $Y = X_3 - n\hat{P}$, der $\hat{P} = \dfrac{X_1 + X_2}{2n}$.

Begrunn at det i vår situasjon er rimelig å anta at Y er tilnærmet normalfordelt, og vis at variansen til Y er $\frac{3}{2}np(1-p)$.

Bruk dette til å utlede et tilnærmet 95% prediksjonsintervall for antallet spurte som i målingen fra Byrå C svarer JA på spørsmålet om regjeringsskifte.

Bestem også intervallet numerisk basert på de observerte verdiene.

Oppgave 3

En fabrikk produserer kabel og tid om annet oppstår det feil på den produserte kabelen. La Z betegne lengden (i kilometer) på kabelen mellom to etterfølgende feil. Vi skal anta at feilene oppstår uavhengig av hverandre, dvs. at påfølgende observasjoner av Z langs kabelen, $Z_1, Z_2, Z_3, \ldots$, er uavhengige stokastiske variabler.

Av erfaring vet en at lengden mellom to etterfølgende feil er eksponensialfordelt med parameter λ, dvs. Z har sannsynlighetstetthet

$$f(z;\lambda) = \begin{cases} \lambda e^{-\lambda z} & \text{for } z > 0, \\ 0 & \text{ellers} \end{cases}$$

og kumulativ fordelingsfunksjon

$$F(z;\lambda) = \begin{cases} 1 - e^{-\lambda z} & \text{for } z > 0, \\ 0 & \text{ellers.} \end{cases}$$

Ved hjelp av fabrikkens opptegnelser over tidligere feil på kabelen ønsker vi å estimere λ. Men det viser seg dessverre at fabrikken ikke har notert nøyaktig lengde på kabelen mellom hver feil, i stedet er det kun notert antall hele kilometer, M, med kabel mellom hver feil. Dvs, dersom $Z < 1.0$ har fabrikken notert seg $M = 0$, dersom $1.0 \leq Z < 2.0$ har fabrikken notert seg $M = 1$, dersom $2.0 \leq Z < 3.0$ har fabrikken notert seg $M = 2$, osv.

a) Vis at punktsannsynligheten for M blir

$$\mathrm{P}(M = m) = (1 - e^{-\lambda})e^{-\lambda m} \quad \text{for} \quad m = 0, 1, 2, \ldots.$$

b) Bestem sannsynlighetsmaksimeringsestimatoren (SME) for λ basert på n observasjoner M_1, $M_2, \ldots$, M_n.

Oppgave 4 Vitamin C

I en medisinsk studie på marsvin ble det benyttet to ulike kilder til vitamin C-inntak. Disse var appelsinjuice (tilskudd 1) og syntetisk askorbinsyre (tilskudd 2). Responsmålet som ble brukt var lengden til odontoblastceller i fortennene til marsvinene. Forskerne hadde som mål å studere effekten av hvert av tilskuddene, og deretter å sammenligne dem.

$X_1, X_2, \ldots, X_{n_1}$ angir odontoblastlengdene til et tilfeldig utvalg av n_1 marsvin som fikk tilskudd 1 og $Y_1, Y_2, \ldots, Y_{n_2}$ angir odontoblastlengdene til et tilfeldig utvalg av n_2 marsvin som fikk tilskudd 2. Vi antar at $X_i \sim N(\mu, \sigma^2)$, dvs. at X_i er normalfordelt med $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$, og at $Y_j \sim N(\eta, \tau^2)$ for $i = 1, 2, \ldots, n_1$ og $j = 1, 2, \ldots, n_2$, og at de to tilfeldige utvalgene er uavhengige.

Totalt fikk $n_1 = 10$ marsvin tilskudd 1 og $n_2 = 10$ marsvin tilskudd 2. Datasettet finner du (sortert) i tabellen under. Lengdene er i mikrometer (10^{-6} meter).

Tilskudd	Observasjoner									
Tilskudd 1: Appelsinjuice	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
Tilskudd 2: Askorbinsyre	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

Deskriptive mål for datasettet er $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 13.18$, $\bar{y} = \frac{1}{10}\sum_{j=1}^{10} y_j = 8.00$, $s_x = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2} = 4.44$, og $s_y = \sqrt{\frac{1}{9}\sum_{j=1}^{10}(y_j - \bar{y})^2} = 2.77$.

a) Først skal vi bare studere situasjonen der appelsinjuice ble gitt, dvs. x-målingene. Hva er en god estimator for forventet lengde på odontoblastcellene, μ? Hva blir estimatet når data er som oppgitt over? Et punktestimat alene forteller ingenting om variablitet, slik at vi også ser på et konfidensintervall for μ. Lag et 90 % konfidensintervall for μ

b) Gjenta det du gjorde i **a)** med askorbinsyre-dataene, dvs. y-malingene, men for η lager du heller et 99 % (og ikke 90) konfidensintervall. Vil generelt et 90% konfidensintervall være smalere eller bredere enn et 99% konfidensintervall (hvis vi laget begge intervallene for η)?

c) Forskerne vil gjerne se på forskjellen i effekten de to tilskuddene har på forventet odontoblastlengde. Definer $\delta = \mu - \eta$ som differansen mellom forventet odontoblastlengde for de to tilskuddene. Hva er en god estimator for δ? Hva blir estimatet når data er som oppgitt over? Skriv til slutt opp formelen for et 95% konfidensintervall for δ og regn ut numerisk verdi. Er det grunn til å tro at de to tilskuddene har ulik effekt på forventet odontoblastlengde? (Dette spørsmålet skal vi jobbe mer med videre i kurset, men du skal her basere svaret ditt på konfidensintervallet du har laget.)

Oppgave 5 Matlaboppgave om konfidensintervall

Hos næringsmiddelprodusenten Lønsjkrønsj fremstilles produktet *Nøttebøtte*, små bøtter som hver inneholder 500 g assorterte nøtter. Ledelsen i Lønsjkrønsj innser at man ikke kan garantere at hver eneste bøtte inneholder nøyaktig 500 g nøtter, men for at det ikke skal oppstå for store avvik, er følgende kvalitetskontrollrutine innført: Ved slutten av hver produksjonsdag velges 12 nøttebøtter tilfeldig fra dagens produksjon. Deretter veies innholdet med en svært nøyaktig vekt, slik at en får observasjonene $X_1, X_2, \ldots, X_{12}$. Anta at observasjonene er uavhengig og identisk normalfordelte med ukjent forventningsverdi μ og ukjent varians σ^2. Anta videre at vekta som brukes er så nøyaktig at vi kan se bort fra måleusikkerheten.

a) Skriv opp uttrykket for et 90% konfidensintervall for μ basert på $X_1, X_2, \ldots, X_{12}$.

Lag en Matlab-funksjon som tar inn en vektor med observasjoner og returnerer nedre og øvre grense til intervallet.

Eksempel på Matlab-kode som kan brukes:

```
function [mu_L, mu_U] = konfint_nb(x)
n = length(x);
alpha = 0.10;
tq = icdf('t', 1-alpha/2, n-1);
xbar = mean(x);
s = std(x);
mu_L = xbar - tq*s/sqrt(n);
mu_U = xbar + tq*s/sqrt(n);
end
```

b) Anta at de sanne parameterverdiene er $\mu = 498.25$ g og $\sigma^2 = 2.67^2$ g^2, og at de holder seg konstante gjennom en periode på 300 produksjonsdager. Hver dag beregnes et nytt konfidensintervall som i **a)**, basert på 12 nye observasjoner.

Hvor mange av de 300 konfidensintervallene forventes å inneholde den sanne verdien av μ?

Lag et Matlab-skript som simulerer de 300 utvalgene fra normalfordelingen med de sanne verdiene av μ og σ^2, og beregner de 300 konfidensintervallene ved å kalle funksjonen fra **a)**.

Hvor mange av konfidensintervallene inneholder den sanne verdien av μ? Kommenter svaret.

Eksempel på Matlab-kode som kan brukes:

```
mu = 498.25; sigma = 2.67;
```

```
antall_innenfor = 0;

for i = 1:300
    x = normrnd(mu, sigma, [1,12]);
    [l,u] = konfint_nb(x);
    if (l <= mu && mu <= u)
        antall_innenfor = antall_innenfor + 1;
    end
end

andel_innenfor = antall_innenfor/300
```

Fasit

1. **a**) 138.89 h **b**) 10 kWh

2. **a**) $\mathrm{E}[\widehat{P}] = p$, $\mathrm{Var}[\widehat{P}] = \frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)$, $\mathrm{E}[P^*] = p$, $\mathrm{Var}[P^*] = \frac{1}{n_1+n_2}p(1-p)$ **b**) $[633, 704]$

3. **b**) $\widehat{\lambda} = \ln\left(1/\overline{M} + 1\right)$ der $\overline{M} = \frac{1}{n}\sum M_i$

4. [10.61,15.75] [5.15,10.85] [1.66,8.70]