Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4240 Statistikk
Høst 2016

Innlevering 3

Dette er den første av to innleveringer i blokk 2. Denne øvingen skal oppsummere pensum forelest i uke 40-43. Øvingen handler om funksjoner av stokastiske variable (kap. 7 i lærebok og notat om ordningsvariable), estimering og konfidensintervall (kap. 8 og 9 i læreboka). Alle deloppgaver teller like mye.

Oppgave 1 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 $\mathrm{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$, $\mathrm{E}(Y_j) = \eta$ og $\mathrm{Var}(Y_j) = \tau^2$ 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 må også se 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 forskjellene de to tilskuddene har på differansen i forventet odontoblastlengde. Definer $\delta = \mu - \eta$ som differansen mellom forventet odontoblastleng-

de 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 2 Surhetsgrad i ferskvann

Man skal undersøke pH-verdien, μ, i et ferskvann. Til dette formål benyttes to ulike målemetoder, metode A og metode B. La X betegne målt verdi ved metode A og la Y tilsvarende betegne målt verdi med metode B. Anta at X og Y er uavhengige og at begge er normalfordelt med forventning μ. Målenøyaktigheten for de to målemetodene er forskjellige. Anta at variansen til X er $\sigma_A^2 = 0.2^2$, mens variansen til Y er $\sigma_B^2 = 0.1^2$. For å estimere μ er det foreslått tre estimatorer

$$\widehat{\mu} = Y \quad , \quad \widetilde{\mu} = \frac{1}{2}X + \frac{1}{2}Y \quad \text{og} \quad \mu^* = \frac{1}{5}X + \frac{4}{5}Y.$$

a) Hvilke to egenskaper kjennetegner en god estimator?

Hvilken av estimatorene $\widehat{\mu}$, $\widetilde{\mu}$ og μ^* vil du foretrekke? Begrunn svaret.

b) Hvilken sannsynlighetsfordeling har μ^*? Begrunn svaret.

Utled et 90%-konfidensintervall for μ ved å ta utgangspunkt i estimatoren μ^*.

Oppgave 3

Ved en redningssentral ankommer alarmer som krever utrykning med redningshelikopter som en poissonprosess. Forventet antall alarmer mottatt i løpet av ett døgn er λ. Vi vet da fra pensum at antall alarmer mottatt i løpet av et døgn er poissonfordelt med forventning λ, og at tiden det går mellom to alarmer er eksponensialfordelt med forventning $1/\lambda$.

a) Anta i dette punktet at det er kjent at $\lambda = 1.5$.

Beregn sannsynligheten for at det kommer to eller flere alarmer i løpet av et døgn.

Beregn sannsynligheten for at det kommer to eller flere alarmer i løpet av et døgn dersom vi vet at det kommer minst én alarm.

Anta at det går 2 timer fra en alarm mottas til redningshelikopteret er klar til ny utrykning. Hva er sannsynligheten for at minst en ny alarm mottas i løpet av denne tiden?

I praksis vil λ være ukjent, men kan estimeres fra observerte data. En mulig fremgangsmåte er å registrere data inntil en har observert et forhåndsbestemt antall, n, alarmer. Observasjonene vil da være $T_1, T_2, \ldots, T_n$, hvor T_1 er tida frem til første alarm, T_2 er tida fra første til andre alarm osv. Disse tidene vil være uavhengig, identisk eksponentialfordelte med samme forventningsverdi $1/\lambda$.

b) Utled sannsynlighetsmaksimeringsestimatoren (SME) $\hat{\lambda}$ for λ basert på $T_1, T_2, \ldots, T_n$. Hva blir estimatet for λ når observasjonene er som gitt ovenfor?

Tabell 1: Observerte alarmtider.

i	1	2	3	4	5	6	7	8	9	10
t_i	1.8	0.2	0.1	1.0	0.3	1.7	0.4	0.8	0.1	1.1

Regn ut forventningsverdien til estimatoren $\hat{\lambda}$. Dersom estimatoren ikke er forventningsrett, foreslå en forventningsrett estimator $\tilde{\lambda}$ basert på $\hat{\lambda}$.

Hint: Du kan bruke, uten å vise det, at

$$E\left(\frac{1}{2\lambda\sum_{i=1}^{n} T_i}\right) = \frac{1}{2(n-1)}.$$

c) Forklar hvorfor det er rimelig å anta at $\sum_{i=1}^{n} T_i = T_1 + T_2 + \cdots + T_n$ er tilnærmet normalfordelt. Bruk dette til å regne ut et tilnærmet 90% konfidensintervall for λ når observasjonene er som gitt i tabell 1.

Oppgave 4 Juletrelysene

Vi ser på en lenke med 36 juletrelys der hvert lys 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 eksponensialfordelt 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 5 Estimatorer for standardavvik

Anta at $X_1, X_2, \ldots, X_n$ er uavhengige, identisk normalfordelte tilfeldige variabler med kjent forventningsverdi $\mu = 0$, og ukjent standardavvik σ. Anta videre at vi har observasjoner $x_1, x_2, \ldots, x_n$ av disse tilfeldige variablene, og at vi ønsker å bruke observasjonene til å estimere σ.

For $i = 1, \ldots, n$ har vi at forventningsverdien til kvadratet av X_i er

$$E(X_i^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2\sigma^2}} dx = \sigma^2,$$

og forventningsverdien til absoluttverdien av X_i er

$$E(|X_i|) = \int_{-\infty}^{\infty} |x| f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |x| e^{-\frac{x^2}{2\sigma^2}} dx = \sqrt{\frac{2}{\pi}}\sigma.$$

For å estimere σ kan vi altså se for oss å bruke estimatorene

$$\hat{\sigma}_1(X_1, \ldots, X_n) = \sqrt{\overline{X^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

og

$$\hat{\sigma}_2(X_1, \ldots, X_n) = \sqrt{\frac{\pi}{2}}\overline{|X|} = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n}\sum_{i=1}^{n} |X_i|.$$

Fordelingene til $\hat{\sigma}_1$ og $\hat{\sigma}_2$ er ikke lette å skrive analytisk. Derfor vil vi bruke stokastisk simulering til å sammenligne de to estimatorene. Oppgaven går altså ut på å undersøke egenskapene til $\hat{\sigma}_1$ og $\hat{\sigma}_2$ ved hjelp av simulering i Matlab.

a) Lag et skript som trekker n realisasjoner fra en $N(0, \sigma^2)$-fordeling, og regner ut de to estimatene $\hat{\sigma}_1$ og $\hat{\sigma}_2$. Test skriptet for $n = 10$ og $n = 100$ med $\sigma = 1$ og $\sigma = 10$. Treffer estimatorene godt?

b) Gjenta prosedyren i **a)** 10 000 ganger. Lagre verdiene av $\hat{\sigma}_1$ og $\hat{\sigma}_2$.

Visualiser resultatene (f.eks. ved å lage histogrammer og boksplott), og bruk figurene, sammen med eventuelle beregninger (finn f.eks. den empiriske variansen til hver estimator), til å avgjøre og argumentere for hvilken av de to estimatorene som er å foretrekke.

Fasit

1. [10.61,15.75] [5.15,10.85] [1.66,8.7]

2. **a**) Foretrekker μ^* **b**) $\mu^* \sim \mathrm{N}(\mu, 0.008)$, $[\mu^* - 0.15, \mu^* + 0.15]$

3. **a**) 0.44, 0.57, 0.12 **b**) $1.33, \mathrm{E}(\widehat{\lambda}) = \lambda n/(n-1), \lambda^* = \widehat{\lambda}(n-1)/n$ **c**) $[0.64, 2.03]$

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