Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4245 Statistikk
Vår 2016

Innlevering 4, blokk II

Dette er den andre og siste innleveringen i blokk 2 og omhander konfidensintervall, hypotesetesting og lineær regresjon.

Oppgave 1

Det vurderes å utbedre en lite trafikkert, men farlig veistrekning i Trondheimsområdet. I den sammenheng blir du bedt om å analysere hvor mye trafikk det er på veien. La X være antall biler som passerer et bestemt punkt på veistrekningen fra kl 16:00 til kl 18:00 på en tilfeldig valgt hverdag. Vi antar at X er poissonfordelt med parameter λ, dvs

$$f(x) = \frac{\lambda^x}{x!} \exp\{-\lambda\}, \quad x = 0, 1, 2, \ldots$$

a) Anta bare i dette punktet at $\lambda = 15$.

Regn ut $\mathrm{P}(X > 20)$ og $\mathrm{P}(10 \leq X < 20)$.

Finn E(X).

Anta at verdien til λ er ukjent i resten av oppgaven. Vi ønsker nå å finne realistiske verdier for λ. Vi observerer derfor antall passerende biler fra kl 16:00 til kl 18:00 på n tilfeldig valgte hverdager, $X_1, X_2, \ldots, X_n$. Vi antar at observasjonene er uavhengige og identisk fordelt med samme poissonfordeling som tidligere beskrevet. Resultatet av målingene for $n = 30$ tilfeldige dager, $x_1, x_2, \ldots, x_{30}$, gir $\sum_{i=1}^{30} x_i = 359$.

b) Definér estimatoren

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Bruk sentralgrenseteoremet til å utlede et tilnærmet $(1 - \alpha) \cdot 100\%$ konfidensintervall for λ basert på estimatoren $\hat{\lambda}$.

Regn ut konfidensintervallet for λ med de gitte dataene og $\alpha = 0.01$.

Kommunen ønsker å samle inn mer data om hvor trafikkert veien er. Derfor registreres det også hvor mange biler som passerer på veien fra kl 18:00 til kl 20:00 på m tilfeldig valgte hverdager, $Y_1, Y_2, \ldots, Y_m$. Vi antar at observasjonene er uavhengige og identisk poissonfordelte med parameter $\lambda/2$, altså halv intensitet i forhold til fra kl 16:00 til kl 18:00. Anta også at $Y_1, Y_2, \ldots, Y_m$ er uavhengige av $X_1, X_2, \ldots, X_n$.

c) Utled sannsynlighetsmaksimeringsestimatoren (maximum likelihood estimatoren) for λ basert på observasjonene $X_1, X_2, \ldots, X_n$ og $Y_1, Y_2, \ldots, Y_m$.

Oppgave 2

Det er i dag svært vanlig at man utfører meningsmålinger for å skaffe seg informasjon om de politiske partiers velgeroppslutning. Dette utføres ved at et utvalg av de stemmeberettigede blir spurt hvilket parti de ville ha stemt på dersom det hadde vært valg den dagen. I denne oppgaven skal vi regne litt på denne situasjonen og vi skal fokusere på oppslutningen til ett bestemt parti, som vi benevner P. For å forenkle situasjonen noe skal vi se bort fra muligheten for at noen ikke vil stemme eller at noen ikke vil svare eller svarer usant når de blir spurt om sitt partivalg.

La N betegne antall stemmeberettigede og la p være andelen av disse N som vil stemme på partiet P. Anta at n personer blir spurt i meningsmålingen og la X betegne antall av disse n som svarer at de ville ha stemt på P. Anta til slutt at de n personene som blir spurt er trukket tilfeldig uten tilbakelegging fra de N stemmeberettigede.

a) Forklar hvorfor X er hypergeometrisk fordelt.

Forklar hvorfor X i denne situasjonen er tilnærmet binomisk fordelt med n forsøk og sannsynlighet p.

Som estimator for p benyttes $\widehat{p} = X/n$.

b) Benytt at X er tilnærmet binomisk fordelt til å bestemme forventning og varians for $\widehat{p}$.

Hvis partiet P har en oppslutning på $p = 0.079$ blant de stemmeberettigede, hvor mange personer, n, må man minst spørre for at standardavviket til $\widehat{p}$ ikke skal overstige 0.010.

Som kjent kan en binomisk fordeling tilnærmes med en normalfordeling dersom np og $n(1-p)$ begge er tilstrekkelig store. I resten av oppgaven kan du anta at dette er oppfylt slik at

$$\frac{X - np}{\sqrt{np(1-p)}}$$

er tilnærmet standard normalfordelt.

La $p_0 = 0.079$ betegne oppslutningen til partiet P ved forrige valg. En ønsker nå å benytte resultatet av meningsmålingen til å undersøke om oppslutningen om P har gått ned siden den gang.

c) Formuler dette som et hypotesetestingsproblem. Velg testobservator og lag en hypotesetest med signifikansnivå $\alpha = 0.05$.

Hva blir konklusjonen på testen dersom man har spurt 1000 personer og 52 av disse svarte at de ville ha stemt på partiet P.

Oppgave 3

En blodgiver ser at hans hemoglobinverdier (X, i g/dl) i perioden 1993 til 2012 tilsynelatende var uavhengige og normalfordelt med forventning 16.14 og standardavvik 0.63.

a) Idrettsutøvere får ikke lov å delta i en konkurranse hvis hemoglobinnivået er over 17.5 (blodet blir for tykt). Hva er sannsynligheten for at blodgiveren ville blitt nektet deltakelse hvis han skulle finne på å melde seg på en konkurranse?

Før større konkurranser, som f.eks. Tour de France på sykkel, blir det tatt hyppige blodprøver av idrettsutøverne for å forhindre doping. Hva er sannsynligheten for at vår blodgiver ville fått minst en av fem målinger over grensa på 17.5? Beskriv kort hvilke forutsetninger som må være oppfylt for at dere skal kunne regne ut denne sannsynligheten.

Utover 2000-tallet fant Blodbanken at den gamle målemetoden hadde for stor usikkerhet, med varians på 0.63^2. En ny målemetode ble derfor introdusert fra og med 2007. Femten verdier til blodgiveren målt med den nye metoden ga $\sum_{i=1}^{15}(x_i - \bar{x})^2 = 2.2$.

b) Formuler en hypotesetest som undersøker om det er grunnlag for å påstå at den nye metoden har lavere varians enn den gamle.

Hvilke antagelser må aksepteres for at testen skal kunne gjennomføres?

Hva blir konklusjonen på testen når dataene er som over, og signifikansnivået er 0.05?

Oppgave 4

En 45-åring startet med løpetrening for 9 år siden, og har hvert år siden deltatt i samme mosjonsløp. Anvendt tid, i minutter, er gitt i tabellen nedenfor.

år i	1	2	3	4	5	6	7	8	9
alder x_i	37	38	39	40	41	42	43	44	45
tid y_i	45.54	41.38	42.50	38.80	41.26	37.20	38.19	38.05	37.45

Det oppgis at $\sum_{i=1}^{9} x_i = 369$, $\sum_{i=1}^{9} y_i = 360.37$, $\sum_{i=1}^{9}(x_i - \bar{x})^2 = 60$, $\sum_{i=1}^{9}(y_i - \bar{y})^2 = 63.28$ og $\sum_{i=1}^{9}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{9}(x_i - \bar{x})y_i = -52.57$.

Vi skal anta at observasjonene kan ses på som realisasjoner av uavhengige normalfordelte variable $Y_1, \ldots, Y_9$, hvor $\mathrm{E}(Y_i) = \alpha + \beta x_i$ og $\mathrm{Var}(Y_i) = \sigma^2$.

a) Skriv opp de vanlige forventningsrette estimatorene $\hat{\alpha}$, $\hat{\beta}$ og $\hat{\sigma}^2$ for α, β og σ^2. Regn ut estimatene for α og β for de gitte dataene.

Plott datasettet og den estimerte regresjonslinjen.

List opp antagelsen som er gjort i regresjonsmodellen, og kommenter om disse ser ut til å være oppfylt her.

Det oppgis at estimatet for σ^2 er 1.568^2.

b) Regn ut et uttrykk for variansen til estimatoren $\hat{\beta}$.

Gjennomfør en test av $H_0: \beta = 0$ mot $H_1: \beta \neq 0$, på signifikansnivå 1%.

Hva blir den praktiske fortolkningen av testen over?

Løperen ønsker å predikere anvendt tid på mosjonsløpet neste gang (alder $x_0 = 46$ år).

c) Regn ut predikert tid.

Det oppgis at $\mathrm{Var}(\hat{\alpha} + \hat{\beta}x_0) = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$. Utled et 95% prediksjonsintervall for Y ved $x_0 = 46$ år. Hva blir intervallet med de oppgitte data?

Hvis løperen ber deg predikere anvendt tid om 15 år (alder 60 år), hva vil du svare da?

Fasit

1. **a**) 0.0830, 0.8051 **b**) [10.34, 13.59]

2. **c**) Forkaster H_0

3. **a**) 0.0154,0.075 **b**) Forkaster H_0

4. **a**) $\widehat{\alpha} = 75.96, \widehat{\beta} = -0.876$ **b**) $\text{Var}(\widehat{\beta}) = \sigma^2 / \sum_{i=1}^{n}(x_i - \overline{x})^2$, Forkast H_0 **c**) $35.66, [31.09, 40.23]$