Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Faglige kontakter under eksamen:
Mette Langaas 988 47 649

EKSAMEN I FAG TMA4245 STATISTIKK

Fredag 6. august 2004

Tid: 09:00–14:00

Tillatte hjelpemidler:
Gult A5 ark med egne håndskrevne notater.
Tabeller og formler i statistikk (Tapir Forlag).
K. Rottmann: *Matematisk formelsamling.*
Kalkulator: HP30S.

BOKMÅL

Sensur: 27. august 2004.

Oppgave 1 Midtveiseksamen

Fra høsten 2004 vil det i emnet "Statistikk" bli innført tellende skriftlig midtveiseksamen. Denne vil bli gitt i form av en flervalgsoppgave ("multiple choice") bestående av $n = 20$ spørsmål som alle har m svaralternativer. Studentene må velge et svaralternativ for hvert spørsmål (det er således ikke lov å svare "blankt" på et spørsmål). For å bestå midtveiseksamen må minst 8 spørsmål være korrekt besvart.

Ole lurer på om han skal la være å lese til midtveiseksamen og heller velge tilfeldige svaralternativer på alle spørsmålene (han vil da ikke engang lese oppgaveteksten før han svarer). Før han bestemmer seg, ber han en studiekamerat regne ut hvor stor sannsynlighet han da har for å bestå midtveiseksamen.

La X være antall korrekte svar Ole får på de $n = 20$ spørsmålene.

a) Forklar hvorfor vi kan anta at X er binomisk fordelt med $n = 20$ og $p = \frac{1}{m}$. (Det er ikke tilstrekkelig å skrive opp de generelle forutsetningene for en binomisk fordeling, betingelsene må relateres direkte til situasjonen som er beskrevet.)

Finn sannsynligheten for at Ole består midtveiseksamen hvis han velger å svare tilfeldig på alle spørsmålene, dvs. $P(X \geq 8)$, når antall svaralternativer er $m = 2$. Finn også $P(X \geq 8)$ for $m = 4$ og $m = 5$.

Hva blir forventet antall korrekte svar, dvs. $E(X)$, når $m = 2, 4, 5$?

Vi benytter videre at hvert spørsmål har $m = 5$ svaralternativer.

Ole bestemmer seg for å bruke midtsemesteruken til å jobbe med statistikk. Dagen før midtveiseksamen setter han seg ned og deler pensum inn i tre kategorier; de delene av pensum han synes han kan godt, de han kan middels godt, og de han kan dårlig.

Vi ser på et tilfeldig valgt spørsmål fra midtveiseksamen og definerer følgende fire hendelser:

G = spørsmålet er fra den delen av pensum som Ole kan godt,
M = spørsmålet er fra den delen av pensum som Ole kan middels godt,
D = spørsmålet er fra den delen av pensum som Ole kan dårlig,
K = Ole svarer korrekt på spørsmålet.

Ole setter opp følgende sannsynligheter:
$P(G) = 0.3$, $P(M) = 0.5$, $P(D) = 0.2$, $P(K|G) = 0.8$, $P(K|M) = 0.4$, $P(K|D) = 0.2$.

b) Vis de fire hendelsene i et venndiagram.

Hva er sannsynligheten for at Ole svarer korrekt på et tilfeldig valgt spørsmål, $P(K)$?

Gitt at Ole svarer korrekt på et tilfeldig valgt spørsmål, hva er da sannsynligheten for at spørsmålet er fra den delen av pensum som Ole kan dårlig, $P(D|K)$?

Oppgave 2 Oljeutslipp

Oljeselskaper som er operatører på norsk kontinentalsokkel må hvert år rapportere forurensende utslipp til myndighetene. Under de fleste reservoarer ligger det en vannsone som etter noen tids produksjon trekkes opp i brønnene og blir en del av "væsken" som produseres på feltene. Vannet skilles fra oljen under produksjonen, men en del olje blir igjen i produksjonsvannet som slippes ut i sjøen.

Myndighetene krever at oljekonsentrasjonen i produksjonsvann som slippes ut i sjøen skal være mindre enn 40 mg/l.

Basert på dagens teknologi kan det antas at oljekonsentrasjonen i et tilfeldig valgt utslipp av produksjonsvann er normalfordelt med forventing $\mu = 21.6$ mg/l og standardavvik $\sigma = 3.4$ mg/l.

a) Finn sannsynligheten for at oljekonsentrasjonen i et tilfeldig valgt utslipp av produksjonsvann er mindre enn 30 mg/l.

Hva er sannsynligheten for at oljekonsentrasjonen i et tilfeldig valgt utslipp overskrider myndighetenes krav?

Finn sannsynligheten for at oljekonsentrasjonen i et tilfeldig valgt utslipp ligger innenfor $[\mu - 3\sigma, \mu + 3\sigma]$.

På en produksjonsplattform tas 24 uavhengige målinger av oljekonsentrasjonen i produksjonsvann i løpet av et døgn.

b) Hva er sannsynligheten for at gjennomsnittet av disse målingene er lavere enn 20 mg/l?

Hva er sannsynligheten for at ingen av de 24 målingene er over 30 mg/l?

I et laboratorium testes det ut en ny rensemetode som fokuserer på å redusere kildene til oljekonsentrasjonen i produksjonsvannet. Den nye rensemetoden krever betydelige investeringer. Før man tar den i bruk ønsker man derfor å være mest mulig sikker på at den virkelig er bedre enn dagens teknologi som altså gir et forventet utslipp på 21.6 mg/l.

Det tas 16 målinger av oljekonsentrasjonen etter bruk av den nye rensemetoden. Disse målingene har et gjennomsnitt på 19.8 mg/l.

Det antas at målingene er uavhengige og normalfordelte med ukjent forventning ν, og med standardavvik τ som i første omgang antas kjent, $\tau = 3.4$ mg/l.

På bakgrunn av de 16 målingene ønsker man nå å beslutte om den nye rensemetoden skal tas i bruk.

c) Formulér dette som et hypotesetestingsproblem ved å definere nullhypotese og alternativ hypotese.

Sett opp en testobservator og bestem forkastningsområdet for testen når man velger signifikansnivå $\alpha = 0.01$. Hva blir konklusjonen?

Anta nå at du får opplyst at man i laboratorieforsøket ikke kjenner standardavviket τ, men at du isteden får vite at det empiriske standardavviket for målingene var 2.6 mg/l. Endrer dette konklusjonen om rensemetoden? Begrunn svaret og gi en kommentar.

d) Utled et 95% konfidensintervall for ν når τ antas ukjent. Regn ut tallsvar for intervallgrensene basert på de gitte verdiene for gjennomsnitt og empirisk standardavvik.

Oppgave 3 Trafikktetthet

Trafikktettheten på en vei skal undersøkes. Tiden mellom to etterfølgende bilpasseringer betegnes X (i sekunder). Når man begrenser seg til vanlige hverdager utenom rushtiden, har man erfaring for at den kumulative fordelingsfunksjonen til X er gitt ved

$$F_X(x) = 1 - \left(1 + \frac{x}{\theta}\right) e^{-\frac{x}{\theta}} \quad \text{for } x > 0, \tag{1}$$

der $\theta > 0$ er en parameter som spesifiserer trafikkmengden på veien som undersøkes.

a) Hva er sannsynligheten for at tiden mellom to etterfølgende biler er høyst θ, dvs. hva er $P(X \leq \theta)$?

Vis at sannsynlighetstettheten til X er gitt ved

$$f_X(x) = \frac{x}{\theta^2} e^{-\frac{x}{\theta}} \quad \text{for } x > 0.$$

La $W = \frac{2X}{\theta}$.

b) Finn den kumulative fordelingsfunksjonen $F_W(w)$ til W.

Vis at W er kjikvadratfordelt med 4 frihetsgrader og bruk dette til å finne forventning og varians for X.

Anta i det følgende at θ er en ukjent parameter. For å få informasjon om θ har man gjort n målinger $X_1, \ldots, X_n$ som antas å være uavhengige med kumulativ fordelingsfunksjon $F_X(x)$.

c) Vis at sannsynlighetsmaksimeringsestimatoren for θ basert på de n målingene er

$$\hat{\theta} = \frac{1}{2n} \sum_{i=1}^{n} X_i$$

Finn forventning og varians for $\hat{\theta}$.

Hvilke krav bør man stille til en god estimator? Er disse tilfredsstilt for $\hat{\theta}$?

d) Man ønsker å teste

$$H_0 : \theta = \theta_0 \text{ mot } H_1 : \theta > \theta_0$$

for en gitt verdi $\theta_0 > 0$.

En statistiker foreslår å bruke testobservatoren

$$Y = \frac{4n\hat{\theta}}{\theta_0}$$

For hvilke verdier av Y skal da H_0 forkastes hvis man ønsker en test med signifikansnivå α?

(*Vink:* Finn fordelingen til Y når H_0 gjelder).