Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Bokmål

Faglig kontakt under eksamen:
Håkon Tjelmeland 7359 3538

EKSAMEN I EMNE ST1201/ST6201 STATISTISKE METODER
Fredag 8. desember 2008
Tid: 09:00–13:00

Hjelpemidler: Kalkulator CITIZEN SR-270X eller HP30S med tomt minne.
Statistiske tabeller og formler, Tapir forlag.
K. Rottman: Matematisk formelsamling.
Ett gult ark (A5 med stempel) med egne formler og notater.

Sensur er ferdig: 29. desember 2008.

Oppgave 1

Anta at $X_1, X_2, \ldots, X_n$ er et tilfeldig utvalg fra en laplacefordeling, dvs. sannsynlighetstettheten for X_i'ene er gitt som

$$f(x) = \frac{1}{2\theta} e^{-\frac{|x|}{\theta}}, \; -\infty < x < \infty.$$

Parameteren $\theta > 0$ er ukjent og skal estimeres.

a) Finn sannsynlighetsmaksimeringsestimatoren, $\widehat{\theta}$, for θ.

Er $\widehat{\theta}$ forventningsrett?

Oppgave 2

Anta at $X_1, X_2, \ldots, X_n$ er et tilfeldig utvalg fra en normalfordeling med forventning μ_x og varians σ^2, og at $Y_1, Y_2, \ldots, Y_m$ er et tilfeldig utvalg fra en normalfordeling med forventning μ_y og varians $k\sigma^2$. Anta dessuten at X_i'ene og Y_i'ene alle er uavhengige av hverandre. Parametrene μ_x og μ_y er ukjente og vi skal i denne oppgaven se på hvordan vi kan estimere og lage konfidensintervall for differansen $\mu_x - \mu_y$.

Som estimator for μ_x og μ_y benyttes henholdsvis

$$\widehat{\mu}_x = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{og} \quad \widehat{\mu}_y = \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

a) Begrunn av $\widehat{\mu}_x - \widehat{\mu}_y$ er normalfordelt.

Vis ved å benytte kjente regneregler for forventningsverdi og varians, samt antagelsene spesifisert over, at

$$\mathrm{E}\left[\widehat{\mu}_x - \widehat{\mu}_y\right] = \mu_x - \mu_y \quad \text{og} \quad \mathrm{Var}\left[\widehat{\mu}_x - \widehat{\mu}_y\right] = \sigma^2\left(\frac{1}{n} + \frac{k}{m}\right).$$

b) Anta i dette punktet at begge parametrene σ^2 og k er kjente. Utled da et $(1-\alpha)\cdot 100\%$-konfidensintervall for differansen $\mu_x - \mu_y$.

I resten av oppgaven skal vi anta at parameteren σ^2 er ukjent, mens parameteren k er kjent.

La S_x^2 og S_y^2 være empirisk varians for henholdsvis X_i'ene og Y_i'ene, dvs.

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \quad \text{og} \quad S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2.$$

Da har vi som kjent at

$$\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{og} \quad \frac{(m-1)S_y^2}{k\sigma^2} \sim \chi^2_{m-1},$$

samt at S_x^2 og $\bar{X}$ er uavhengige stokastiske variabler og at S_y^2 og $\bar{Y}$ er uavhengige stokastiske variabler.

c) Vis at

$$S_p^2 = \frac{n-1}{n+m-2}S_x^2 + \frac{m-1}{k(n+m-2)}S_y^2$$

er en forventningsrett estimator for σ^2 og at

$$\frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Vis videre at

$$T = \frac{(\widehat{\mu}_x - \widehat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{k}{m}\right)}}$$

er Student t-fordelt med $n+m-2$ frihetsgrader. I utledningene i dette punktet, spesifiser spesielt hvilke kjente egenskaper og sammenhenger mellom ulike typer fordelinger du benytter, og forklar hvorfor vilkårene for disse egenskapene er oppfylt i den aktuelle situasjonen.

d) Utled et $(1-\alpha)\cdot 100\%$-konfidensintervall for differansen $\mu_x - \mu_y$.

Anta at du får lov til å være med i planleggingen av forsøket. Totalt antall målinger som kan utføres, $N = n + m$, er gitt ut fra økonomien i prosjektet, men under denne bibetingelsen kan du bestemme antall målinger av hver type, n og m. Hvordan vil du velge n og m for at forventet lengde på konfidensintervallet skal bli minst mulig?

Oppgave 3

La X være inntekten til en tilfeldig valgt lønnsmottager i ei bestemt befolkningsgruppe. Det er da mye vanlig å anta at X er paretofordelt, dvs at X har sannsynlighetstettheten

$$f(x) = \begin{cases} \frac{\theta c^\theta}{x^{\theta+1}} & \text{når } c < x < \infty, \\ 0 & \text{ellers,} \end{cases}$$

hvor c er minsteinntekten i denne befolkningsgruppa og $\theta > 1$ er en parameter som beskriver lønnsforskjellene i gruppa. I denne oppgaven skal vi forutsette at minsteinntekten c er kjent, mens θ er ukjent og skal estimeres fra observerte data.

a) La $Y = 2\theta(\ln X - \ln c)$. Vis at Y er kji-kvadrat fordelt med to frihetsgrader.

For å estimere parameteren θ gjør man n observasjoner $X_1, X_2, \ldots, X_n$ av inntekter i den aktuelle befolkningsgruppa. Vi skal anta at disse utgjør et tilfeldig utvalg fra paretofordelingen gitt over. Sannsynlighetsmaksimeringsestimatoren (SME) for θ blir da (du trenger **ikke** å vise dette)

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln X_i - n \ln c}.$$

Vi ønsker så å benytte observasjonene $X_1, X_2, \ldots, x_n$ til å bestemme om det er grunnlag for å påstå at $\theta > \theta_0$, der θ_0 er et gitt tall. Vi har altså en hypotesetestingssituasjon med hypotesene

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

Anta at vi vil benytte beslutningsregelen at man skal forkaste H_0 dersom $\widehat{\theta} \geq k$ der k bestemmes slik at sannsynligheten for type I feil er lik et gitt signifikansnivå α, dvs

$$\mathrm{P}\left(\widehat{\theta} \geq k \,\middle|\, H_0 \text{ er riktig}\right) = \alpha.$$

b) Begrunn at $2n\theta/\widehat{\theta}$ er kji-kvadrat fordelt med $2n$ frihetsgrader.

Benytt dette til å utlede en formel for k som funksjon av n, θ_0 og α. Beregn deretter numerisk verdi for k når $n = 20$, $\theta_0 = 2.0$ og $\alpha = 0.01$.

c) Når $n = 20$, $\theta_0 = 2.0$ og $\alpha = 0.01$, bestem for hvilken verdi av θ sannsynligheten for type II feil blir lik 0.05.