UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Fasit - Statistiske metoder og dataanalyse 1.
Eksamensdag:	Torsdag 12. desember 2013.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over normal-, t-, og χ^2-fordeling.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110.

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

Oppgave 1.

Har $X_1, X_2, ..., X_m$ u.i.f. N(μ, σ_1^2) og $Y_1, Y_2, ..., Y_n$ u.i.f. N(μ, σ_2^2). Målingene fra instrument A og B er også uavhengige av hverandre. De to variansene σ_1^2 og σ_2^2 er kjent for de to instrumentene.

a) Start med at $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ er N$(\mu, \sigma_1^2/m)$. Derfor er $(\bar{X} - \mu)\sqrt{m}/\sigma_1 \sim$N(0,1) og vi kan sette opp

$$P(-z_{\alpha/2} \leq \frac{(\bar{X} - \mu)}{\sigma_1/\sqrt{m}} \leq z_{\alpha/2}) = 1 - \alpha.$$

Rydder opp slik at vi får intervallet $\bar{x} \pm z_{\alpha/2}\sigma_1/\sqrt{m}$. For 99% ($\alpha = 0.01$) konfidensintervall bruker vi $z_{\alpha/2} = z_{0.005} = 2.58$.

Tolkning: hvis vi konstruerer veldig mange konfidensintervall på denne måten, vil intervallene i det lange løp inneholde den sanne ukjente μ 99% av gangene.

b) *Lengde* $= 2z_{\alpha/2}\sigma_1/\sqrt{m} \leq w$ gir at $m \geq (2z_{\alpha/2}\sigma_1/w)^2$. Hvis $w = \sigma_1$ blir $m \geq (2z_{\alpha/2})^2$, og innsatt $z_{\alpha/2} = z_{0.005} = 2.58$ gir det $m \geq 27$.

c)

$$L(x_1, ..., x_m, y_1, ..., y_n, \mu, \sigma_1^2, \sigma_2^2) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-\frac{1}{2\sigma_1^2}(x_i - \mu)^2) \cdot \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-\frac{1}{2\sigma_2^2}(y_i - \mu)^2)$$

$$\log L \propto -\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i - \mu)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

(Fortsettes på side 2.)

Deriver $\log L$ mhp. μ og sett lik 0. Løser ut μ som

$$\mu = \frac{\frac{\sum x_i}{\sigma_1^2} + \frac{\sum y_i}{\sigma_2^2}}{\frac{m}{\sigma_1^2} + \frac{n}{\sigma_2^2}}.$$

Maximum Likelihood estimator blir da

$$\hat{\mu}_{ML} = \frac{\frac{m\bar{X}}{\sigma_1^2} + \frac{n\bar{Y}}{\sigma_2^2}}{\frac{m}{\sigma_1^2} + \frac{n}{\sigma_2^2}}.$$

d) $\hat{\mu}_{ML}$ er et vektet gjennomsnitt av de to gjennomsnittene. Gjennomsnitt basert på flere observasjoner og fra instrument med lav varians får mer vekt, hvilket er rimelig. Enkelt å vise forventningsrett siden $E(\bar{X}) = E(\bar{Y}) = \mu$. Enkel regning gir

$$V(\hat{\mu}_{ML}) = \frac{1}{\frac{m}{\sigma_1^2} + \frac{n}{\sigma_2^2}}.$$

Oppgave 2.

a) Modell med standard antakelser side 604 i boken.

$$\mathrm{H}_0 : \beta_1 = 0 \quad \text{mot} \quad \mathrm{H}_a : \beta_1 \neq 0.$$

Testobservator

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

der $\hat{\beta}_1$ er estimert effekt av gen 1 og $s_{\hat{\beta}_1}$ er standard feil. Er t-fordelt med 82 frihetsgrader når nullhypotesen er sann. Forkaster H_0 når t er meget stor eller meget liten. Finner forkastningsområde fra krav om $P(typeIfeil) = P(forkast\mathrm{H}_0|\mathrm{H}_0riktig) \leq \alpha = 0.05$. Gir 'Forkast H_0 hvis $t \geq t_{0.025,82}$ eller $t \leq -t_{0.025,82}$'.
Finner ikke 82 frihetsgrader i tabellen, sikrer oss ved å bruke 60, da skal vi forkaste hvis $t \geq 2$ eller $t \leq -2$. Output i (I) gir $t = \frac{-1.1297}{0.3609} = -3.13$ og vi forkaster.

b) P-verdi er sannsynligheten for å få en observator som er minst så ekstrem i forhold til nullhypotesen som det vi faktisk har observert, gitt at nullhypotesen er sann. Evt. minste signifikansnivå som ville ha gitt forkastning.
P-verdien $= 2P(T \geq 3.13)$ der T er t-fordelt m 82 frihetsgrader. Fra tabell finner vi at $2P(T \geq 3.13) \leq 2 \cdot 0.005 = 0.01$ (60 frihetsgrader).

(Fortsettes på side 3.)

c) 95% konfidensintervall for β_1: $\hat{\beta}_1 \pm t_{0.025,82} s_{\hat{\beta}_1}$. Bruker 60 frihetsgrader og finner intervallet (-1.8515, -0.4079). En tosidig test med nivå α for en parameter kan utføres ved å sjekke om den hypotiserte verdien ligger utenfor et $(1-\alpha)100\%$ konfidensintervall for parameteren. Ser her at 0 ikke ligger i 95% konfidensintervall, som stemmer med at vi fikk forkastning på nivå $\alpha = 0.05$.

d) I resultat (II) ser vi at gen 1 ikke lenger er signifikant når vi inkluderer gen 2. Gene 2 er meget signifikant. Hvis vi har informasjon om gen 2, inneholder ikke gen 1 nok tilleggsinformasjon om bentettheten til at den blir signifikant. Dette kan skje når de to kovariatene er korrelert. Vi ser at korrelasjonen mellom gen 1 og gen 2 er 0.73.

e) For den andre modellen er alle kovariater signifikante, og Adjusted R^2 er litt høyere, slik at denne regnes som bedre enn den første. For den andre modellen tolker vi de estimerte effektene slik: Høy genekspresjon for gen 2 har en tendens til å redusere forventet bentetthet, når de andre kovariatene (genene) holdes uendret. Høy genekspresjon for gen 3 har en tendens til å øke forventet bentetthet, når de andre kovariatene (genene) holdes uendret. Samme for gen 4. Vi ser også at gen 3 har størst estimert effekt.

Oppgave 3.

a) Dette er en standard large sample test for en andel, kap. 9.3 i boken. Alternativ hypotese at $p > 0.25$. Finner at testobservator er $z = 2.21$ og P-verdi $= P(Z \geq 2.21) = 0.0136$. Forkaster med nivå 0.05, forkaster ikke med nivå 0.01. Ikke klart at studentene er mer deprimerte enn resten av befolkningen.

Oppgave 4.

X har sannsynlighetstetthet $f(x) = \frac{\theta c^\theta}{x^{\theta+1}}$ når $c < x < \infty$, og $f(x) = 0$ ellers. Her er c minsteinntekten i befolkningsgruppen, mens $\theta > 0$ er en parameter som beskriver lønnsforskjellene i befolkningsgruppen. Har c kjent, mens θ skal estimeres fra uavhengige observasjoner $X_1, X_2, ..., X_n$. Har

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n \log X_i - n \log c}$$

og at $2n\theta/\hat{\theta}$ er χ^2-fordelt med $2n$ frihetsgrader. Skal teste

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{mot} \quad \mathrm{H}_a : \theta > \theta_0.$$

a) Forkaster når $\hat{\theta} \geq k$, der k finnes fra

$$P(\hat{\theta} \geq k | \theta = \theta_0) \leq \alpha$$

$$P(\frac{1}{\hat{\theta}} \leq \frac{1}{k} | \theta = \theta_0) \leq \alpha$$

(Fortsettes på side 4.)

$$P(\frac{2n\theta_0}{\hat{\theta}} \leq \frac{2n\theta_0}{k}) \leq \alpha$$

der $2n\theta_0/\hat{\theta}$ er χ^2-fordelt med $2n$ frihetsgrader under H_0. Da må

$$\frac{2n\theta_0}{k} \leq \chi^2_{1-\alpha,2n}$$

og vi får $k = 2n\theta_0/\chi^2_{1-\alpha,2n}$

b) Innsatt får vi 'Forkast H_0 dersom $\hat{\theta} \geq 3.61'$. Finner $\hat{\theta} = 4.22$ og forkaster H_0 på nivå 0.01.

c)

$$\beta(\theta') = P(\hat{\theta} \leq 3.61|\theta = \theta')$$
$$= P(\frac{2n\theta}{\hat{\theta}} \geq \frac{2n\theta}{3.61})|\theta = \theta')$$
$$= P(\chi^2 \geq \frac{2 \cdot 20 \cdot \theta'}{3.61})$$

der χ^2 er χ^2-fordelt med 40 frihetsgrader. Krever $\beta(\theta') \leq 0.05$

$$P(\chi^2 \geq 11.1 \cdot \theta') \leq 0.05$$

$$11.1 \cdot \theta' \geq 55.76$$

$$\theta' \geq 5.02$$