UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — KORT FASIT.
Eksamensdag:	Mandag 1. desember 2014.
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.

$X_1, X_2, ..., X_n$ er uavhengige Poissonfordelte variable med parameter λt_i.

a)

$$E(\hat{\lambda}) = \frac{\sum_{i=1}^n E(X_i)}{\sum_{i=1}^n t_i} = \frac{\sum_{i=1}^n \lambda t_i}{\sum_{i=1}^n t_i} = \frac{\lambda \sum_{i=1}^n t_i}{\sum_{i=1}^n t_i} = \lambda$$

$$V(\hat{\lambda}) = \frac{\sum_{i=1}^n V(X_i)}{(\sum_{i=1}^n t_i)^2} = \frac{\sum_{i=1}^n \lambda t_i}{(\sum_{i=1}^n t_i)^2} = \frac{\lambda \sum_{i=1}^n t_i}{(\sum_{i=1}^n t_i)^2} = \frac{\lambda}{\sum_{i=1}^n t_i}$$

Innsatt data: $\hat{\lambda} = 3 \cdot 10^{-4}$.

b)

$$L(x_1, \cdots, x_n, t_1, \cdots, t_n; \lambda) = \prod_{i=1}^n \frac{(\lambda t_i)^{x_i}}{x_i!} \exp(-\lambda t_i)$$

$$\log L = \sum_{i=1}^n x_i \log \lambda + \sum_{i=1}^n x_i \log t_i - \sum_{i=1}^n \log(x_i!) - \lambda \sum_{i=1}^n t_i$$

Deriver $\log L$ mhp. λ og sett lik 0:

$$\frac{\partial \log L}{\partial \lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - \sum_{i=1}^n t_i = 0$$

Løser ut λ som

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n t_i}$$

(Fortsettes på side 2.)

og får ML-estimatoren

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

c) En annen mulig estimator for λ er

$$\hat{\hat{\lambda}} = \frac{\sum_{i=1}^n X_i t_i}{\sum_{i=1}^n t_i^2}.$$

Denne er også forventningsrett (bør vises (enkelt)) og har varians

$$V(\hat{\hat{\lambda}}) = \frac{\lambda \sum_{i=1}^n t_i^3}{(\sum_{i=1}^n t_i^2)^2}.$$

Kan vise ved innsetting at $V(\hat{\lambda}) \leq V(\hat{\hat{\lambda}})$ og at vi dermed skal foretrekke $\hat{\lambda}$. Bør referere til MVUE og ML-estimatorenes fordelaktige egenskaper, se s. 357 i boken. Alternativt må man vise generelt at

$$\frac{1}{\sum_{i=1}^n t_i} \leq \frac{\sum_{i=1}^n t_i^3}{(\sum_{i=1}^n t_i^2)^2}$$

dvs.

$$(\sum_{i=1}^n t_i^2)^2 \leq \sum_{i=1}^n t_i^3 \sum_{i=1}^n t_i.$$

(For $t_i = t \; \forall i$, får vi likhet.)

Oppgave 2.

a) Vi gjør en standard large sample hypotesetest for en andel p, se kap. 9.3. Hypotesene blir

$$\mathrm{H}_0 : p = 0.4 \quad \text{mot} \quad \mathrm{H}_a : p > 0.4$$

Vi har $\hat{p} = 0.422$ og $n = 960$. Finner at testobservator under H_0 er $z = 1.391$ og P-verdi $= P(Z \geq 1.391) = 0.082$. Med nivå 0.05 har vi ingen grunn til å forkaste H_0. Vi kan derfor ikke konkludere med at Ap ville ha fått mer enn 40% av stemmene den dagen.

Oppgave 3.

a)
Vi har $X_i \sim N(\mu_A, \sigma^2)$, $i = 1, ..., n$ og $Y_i \sim N(\mu_B, \sigma^2)$, $i = 1, ..., m$. Hypotesene blir

$$H_0 : \mu_A - \mu_B = 0 \qquad H_a : \mu_A - \mu_B > 0$$

Baserer oss på $\bar{X} \sim N(\mu_A, \sigma^2/n)$ og $\bar{Y} \sim N(\mu_B, \sigma^2/m)$. Med kjent varians kan vi bruke en Z-test. Vi skal forkaste nullhypotesen på nivå α når testobservator $Z =$

(Fortsettes på side 3.)

$(\bar{X}-\bar{Y})/\sigma\sqrt{1/n+1/m}$ er større enn z_α, fordi $Z \sim N(0,1)$ under H_0. Med signifikansnivå $\alpha = 0.01$ blir forkastningsområdet verdier over $z_{0.01} = 2.326$. Gitt $\sigma^2 = 4$, $n = 4$, $m = 5$, $\bar{x} = 13$ og $\bar{y} = 10$, blir observert testobservator $z = 2.236$. Nullhypotesen kan derfor ikke forkastes. Data gir ikke grunnlag for å påstå at strand A er mer forurenset enn strand B på nivå 0.01.

b) P-verdi er sannsynligheten for å observere det man har observert, eller noe mer ekstremt, gitt at H_0 er riktig. Evt. minste signifikansnivå som hadde gitt forkastning med de observasjonene man har. Her: Sannsynligheten for å observere en forskjell i gjennomsnitt på +3 eller mer, med 4 og 5 observasjoner, gitt at det egentlig ikke er noen forskjell i forventet forurensningsnivå. P-verdi = $P(Z \geq 2.236) = 0.0125$.

c) Setter $m = n$. Krever

$$P(Z \geq 2.326 \,|\, \mu_A - \mu_B = 4) \geq 0.95$$

der $Z = (\bar{X}-\bar{Y})/\sigma\sqrt{1/n+1/n}$. Nå er forventningen til $\bar{X}-\bar{Y}$ ikke lenger 0, men 4, så vi må omordne slik at vi får

$$P(\frac{\bar{X}-\bar{Y}-4}{\sigma\sqrt{1/n+1/n}} \geq 2.326 - \frac{4}{\sigma\sqrt{1/n+1/n}}) \geq 0.95,$$

der uttrykket til venstre nå er standard normalfordelt. Dette gir at uttrykket til høyre må være mindre eller lik $-z_{0.05}$ og vi finner $n \geq 7.88$, dvs. vi må ha minst 8 observasjoner fra hver strand.

d) Fordi vi antar lik varians bruker vi $S_p^2 = \frac{(n-1)S_A^2+(m-1)S_B^2}{n+m-2}$ som estimator for σ^2. Innsatt finner vi da estimatet $S_p^2 = 7.143$.
Testobservator

$$t = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{1/n+1/m}} \sim t_{n+m-2},$$

er student-t-fordelt med $n+m-2$ frihetsgrader. Med $n = 4$ og $m = 5$ blir dette 7 frihetsgrader. Observert verdi for testobservator blir $t = 1.67$. Testen blir at vi skal forkaste H_0 for verdier større enn $t_{0.01,7} = 2.998$. Med nivå $\alpha = 0.01$ har vi ingen grunn til å forkaste H_0.

e) Standard to-utvalgs-t-intervall, $3 \pm 3.499 \cdot \sqrt{7.143} \cdot \sqrt{1/4+1/5}$, gir intervallet (-3.273, 9.273). Tolkning se bok kap. 8.1.

Oppgave 4.

a) $\hat{\beta}_0 = 53.33$, $\hat{\beta}_1 = 0.533$ og $\hat{\sigma} = 1.035$. Må kommentere: 1) scatterplott med tilpasset linje, god tilpasning 2) R-squared=0.9887, stor forklaringsgrad 3) alle regresjonskoeffisienter signifikante! 4) qq-plott: normalitetsantakelsen ok 5) residualplottet: ikke noe mønster, feks. konstant varians ok.

(Fortsettes på side 4.)

b) Har at $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$er t-fordelt med 16 fr.g. Standard utledning gir at intervallet blir $\hat{\beta}_1 \pm t_{0.025,16} s_{\hat{\beta}_1}$. Fra utskriften har vi $\hat{\beta}_1 = 0.533$ og $s_{\hat{\beta}_1} = 0.01428$. Bruker $t_{0.025,16} = 2.12$. Gir 95% konfidentintervall som (0.5027,0.5633).

c) Ved å konstantere at 0 ikke dekkes av 95% konfidensintervallet ovenfor, vet vi at vi kan forkaste nullhypotesen $\mathrm{H}_0 : \beta_1 = 0$ til fordel for den tosidige alternativhypotesen $\mathrm{H}_a : \beta_1 \neq 0$ på signifikansnivå $\alpha = 0.05$. Generelt har vi sammenheng mellom tosidig testing på nivå α og et $(1-\alpha)\cdot 100\%$ konfidensintervall.

d) Konfidensintervall for forventet løselighet $\mu_{Y \cdot x^*}$ i punktet x^*, forklaring s. 656 i boken. Prediksjonsintervall for en ny observasjon Y i punktet x^*, forklaring s. 659. Pluss for å utlede formlene og tegne. For $x^* = 35$ grader Celcius finner vi konfidensintervallet (71.47, 72.51) og prediksjonsintervallet (69.74,74.24) fra utskriften.

e)

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

der ϵ_i er iid. $N(0, \sigma^2)$, $i = 1, ..., n$. Ser at 1) annengradsleddet er ikke signifikant og 2) Adjusted R-squared er ikke større enn for den lineære modellen. Den lineære modellen så også ut til å passe godt. Ingen grunn til å bruke et annengradsledd for disse dataene, for de gitte temperaturene. Mulig denne effekten ville blitt synlig for andre temperaturer.