UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk - kortfattet fasit
Eksamensdag:	Mandag 3. desember 2018.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 2 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Lærebok Moore, McCabe & Craig: Introduction to the practice of statistics. Det er tillatt å notere i læreboka, men ikke lov å klistre inn notater. Ordliste. Kapittel 14, som deles ut. Godkjent kalkulator.

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

Det er fire oppgaver med til sammen ti del-spørsmål. Hvert del-spørsmål teller likt.

Oppgave 1.

a) $P(X > 1) = 1 - P(X \leq 1) = 1 - P(Z \leq \frac{1-0}{0.5}) = 1 - 0.9772 = 0.0228$.

b) Må først komme frem til at $W \sim N(\mu, \sqrt{\sigma_X^2 + \sigma_Y^2})$, innsatt $W \sim N(2, 1.58)$. Finner så $P(W > 3) = 1 - P(W \leq 3) = 1 - P(Z \leq \frac{3-2}{1.58}) = 1 - 0.7357 = 0.2643$.

Oppgave 2.

a) Antar at vi har to uavhengige tilfeldige utvalg, og at de $n_1 = 21$ uavhengige observasjonene x_{1i} fra mais-populasjonen kommer fra en normalfordeling med forventning μ_1 og at de $n_2 = 24$ uavhengige observasjonene x_{2j} fra bomulls-populasjonen kommer fra en normalfordeling med forventning μ_2.
Tester hypotesene $H_0 : \mu_1 - \mu_2 = 0$ mot $H_a : \mu_1 - \mu_2 \neq 0$
Siden den empiriske variansen til bomulls-utvalget er ca. 14 ganger større enn variansen til mais-utvalget, er det ikke naturlig å anta lik varians. Vi bruker derfor en to-utvalgs-t-observator på formen

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

der $\bar{x}_1$ og $\bar{x}_2$ er gjennomsnittet i de to gruppene, og s_1^2 og s_2^2 er de to empiriske

(Fortsettes på side 2.)

variansene. Når H_0 er riktig, er denne test-observatoren tilnærmet t-fordelt, med tilnærmet k frihetsgrader, der k finnes fra programvare (R).

b) Vi bruker den første t-testen, som ikke antar lik varians. Antall frihetsgrader er beregnet til 26.57. Verdien på t-observatoren ovenfor er 8.3477, og tilhørende p-verdi for hypotesene ovenfor er $6.67 \cdot 10^{-9}$. Vi forkaster H_0 og konkluderer med at det er forskjell i avstanden mellom øynene for de to diettene, på ethvert rimelig signifikansnivå.

c) Et 95% konfidensintervall for forskjellen finnes ved $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{s_1^2/n_1 + s_2^2/n_2}$, der $t^* = 2.056$ hvis vi bruker tilnærmet 26 frihetsgrader (fra R). Gir konfidensintervallet (0.3800,0.6284). Dvs. større avstand mellom øynene for fluene som ble alet opp på mais.

Oppgave 3. Estimert odds-ratio er $OR = e^{0.78} = 2.18$, og vi kan konkludere med at en enhets økning i x omtrent dobler oddsen for å finne mineralet. Konfidensintervallet for odds-ratio e^{β_1} finnes ved å transformere et konfidensintervall for β_1, og vi får formelen $(e^{b_1 - z^* SE_{b_1}}, e^{b_1 + z^* SE_{b_1}})$. For et 95% konfidensintervall har vi $z^* = 1.96$ og intervallet blir (1.83, 2.60).

Oppgave 4.

a) Korrelasjonen finnes som roten av Multiple R-squared i utskriften, dvs. 0.776. Positiv fordi b_1 er positiv. $b_0 = 0.175$, $b_1 = 0.00319$. b_1 er estimert endring i forventet gjennomsnittlig is-konsum per hode for en økning på 1 grad Celsius i ukes-temperaturen. $s = 0.0241$.

b) Alle koeffisienter er signifikante, normalfordelingsplottet stemmer rimelig godt med normalfordelingsantakelsen, det er ingen mønster i residualplottet, så antakelser om linearitet og konstant varians OK. Bør begrunne kort hvorfor plottene er ok (kommentere mulig outlier). Modellen passer rimelig bra med dataene. At $R^2 = 0.6016$ betyr at ca. 60% av variasjonen i responsvariabelen (gj. snittlig isforbruk per hode) kan forklares av den lineære sammenhengen med forklaringsvariabelen (ukestemperatur).

c) $SE_{b_1} = 0.00319/6.502 = 0.00049$. 28 frihetsgrader gir konfidensintervallet (0.00219, 0.00419) for β_1. Siden 0 ikke er i 95%-konfidensintervallet, kan vi forkaste $H_0 : \beta_1 = 0$ og konkludere med den tosidige $H_a : \beta_1 \neq 0$ på signifikansnivå 0.05 (stemmer med p-verdien i utskriften).

d) Fra utskriften finner vi prediksjonsintervallet (0.2075,0.3138) liter i uken (95% er default). Men $x = 27$ ligger klart utenfor det området vi har data for. Det er ikke sikkert at vi vil ha en lineær sammenheng for høye temperaturer. Noe tvilsom ekstrapolasjon.

SLUTT