UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Onsdag 3. desember 2014
Tid for eksamen:	09.00 – 13.00
Oppgavesettet er på 3 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Lærebok: Moore, McCabe & Craig: Introduction to the practice of statistics, ordliste for bruk i STK1000, godkjent kalkulator.

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

Løsningsforslag

Oppgave 1

a) Forsøksplanen kalles parplan. På engelsk "matched pair design".

Ideen er å dele enhetene i undersøkelsen inn i par som er så like som mulig etter bakgrunnsvariable. Effekten av to behandlinger, hvorav en kan være kontroll (placebo), sammenlignes innen hvert par. Ideelt skal randomisering brukes for å bestemme hvilken enhet innen hvert par som skal ha behandling eller kontroll.

I dette tilfelle er enhetene personene målt på to tidspunkter, og parene representerer hvert enkelt individ. Fysioterapi er behandling og kontroll er tilstanden før behandling.

Her er behandling og kontroll gitt, og kan ikke randomiseres som tilfellet er hvis behandling er en ny medisin, og kontroll er en narremedisin eller placebo.

Et annet spørsmål er utvelgelsen av enhetene som inngår i undersøkelsen, dvs utvalgsplanen. For å unngå skjevhet bør dette ideelt gjøres ved bruk av sannsynlighetsutvalg. Om det er mulig, vil avhenge av omstendighetene rundt undersøkelsen, for eksempel hvor vanlig denne typen kneoperasjoner er.

b) Gjennomsnitt: $\bar{x}$= (-1-6$\cdots$ -8)= -24/7= -3.43
Sum av kvadratavvik: $(-1 - (-3.428571))^2 + \cdots + (-8 - (-3.428571))^2 = 73.71429$
Estimat for varians: $s^2 = 73.71429/6 = 12.28571$
Estimat for standardavvik: $s = \sqrt{12.28571} = 3.51$
Standardfeil for gjennomsnitt: $s/\sqrt{n} = 3.505098/\sqrt{7} = 1.32$

(Fortsettes på side 2.)

c) Hvis de n=7 differansene er $N(\mu,\sigma)$-fordelt, er $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ t-fordelt med n-1 frihetsgrader, og et 95% konfidensintervall har grenser

$$\bar{x} \pm t^* s/\sqrt{n}$$

der $P(T \geq t^*) = 0.025$ for en t(n-1) fordelt tilfeldig variabel.

I dette tilfellet er n=7 slik at $0.025 = P(T \geq 2.447)$ og grensene

$$-3.43 \pm 2.447 \times 3.51/\sqrt{7} = -3.43 \pm 2.447 \times 1.32 = \begin{cases} -6.68 \\ -0.18 \end{cases}$$

Forutsetningen for at konfidensintervallet skal ha konfidensgrad nøyaktig 95% er at differansene kan oppfattes som et enkelt tilfeldig utvalg fra en populasjon av variable som er $N(\mu,\sigma)$-fordelte. Det medfører at det standariserte gjennomsnittet er t(n-1)-fordelt.

d) Konfidensintervallet fra punkt c) inneholder ikke 0. Det betyr den tosidige testen for null hypotesen om at forventningen til differansene er lik 0 må forkastes med signifikans nivå 5%.

Her er testobservatoren $t = \frac{-3.43}{1.32} = -2.60$. Siden testen er tosidig er P-verdien $P = 2P(T \geq 2.60)$. Men fra tabell er $P(T \geq 2.447) = 0.025$. og $P(T \geq 2.612) = 0.020$, slik at $0.004 = 2 \times 0.002 < P < 2 \times 0.025 = 0.005$. P-verdien er derfor mellom 4% og 5%.

e) En ny observasjon har formen $y_8 = \mu + \epsilon_8$ der $\mu_{\bar{x}} = \mu$ og ϵ_8 er $N(0,\sigma)$-fordelt. Siden forventningen til ϵ_8, $\mu_{\epsilon_8} = 0$, er $\bar{x}$ en forventningsrett estimator for y_8. Det betyr at en estimator for y_8 vil ha varians $\sigma^2_{y_8} = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2(1 + \frac{1}{n})$, så standardavviket estimeres med $s\sqrt{(1+\frac{1}{n})}$.

Under antagelsen om at differansene kan oppfattes som et enkelt tilfeldig utvalg fra en populasjon av variable som er $N(\mu,\sigma)$-fordelte, vil den standarisret variabelen være t(n-1), n=7, fordelt. Et 95% prediksjonsintervall her derfor formen

$$\bar{x} \pm t^* s\sqrt{1+\frac{1}{n}}$$

der t*=2.447 som tidligere. I dette tilfellet er grensene til prediksjonsintervallet

$$-3.43 \pm 2.447 \times 3.51\sqrt{1+\frac{1}{7}} = -3.43 \pm 9.18 = \begin{cases} -12.61 \\ 5.75 \end{cases}$$

Oppgave 2

a) $x^* = 15$ gir predikert verdi $\hat{y} = b_0 + b_1 x^* = 13.49 + 1.065 \times 15 = 29.47$.

b) Under antagelsen om at variablene er normalfordelt er $\frac{b_1-\beta_1}{SE_{b_1}}$ t-fordelt med n-2= 10-2=8 frihetsgrader. Et 95% konfidensintervall har grenser

$$b_1 \pm t^* SE_{b_1}$$

(Fortsettes på side 3.)

der $P(T \geq t^*) = 0.025$ for en t(n-2) fordelt tilfeldig variabel T. Men $P(T \geq 2.306) = 0.025$ slik at grensene er

$$1.065 \pm 2.306 \times 0.218 = \begin{cases} 0.56 \\ 1.57 \end{cases}$$

c) Tester $H_0 : \beta_1 = 0$ mot $H_a : \beta_1 > 0$. Testobservator: $t = \frac{b_1}{SE_{b_1}} = 4.89$.

Da er P-verdien $P = P(T \geq 4.89)$ for en t(8)-fordelt variabel T. Fra tabellen $P(T \geq 4.501) = 0.001$ og $P(T \geq 5.041) = 0.0005$, slik at P-verdien ligger mellom 0.0005 og 0.001, nærmest 0.0005.

Oppgave 3

a) Hvis X er Bin(n,p) fordelt kan X representeres som $X = S_1 + \cdots S_n$ der $S_1, \ldots, S_n$ er uavhegige og $P(S_i = 0) = 1-p$ og $P(S_i = 1) = p, i = 1, \ldots, n$.

Da er $\mu_{S_i} = 0 \times (1-p) + 1 \times p = p$ og $\sigma^2_{S_i} = (0-\mu_{S_i})^2 P(S_i = 0) + (1-\mu_{S_i})^2 P(S_i = 1) = p^2(1-p) + (1-p)^2 p = (1-p)p(p+1-p) = p(1-p)$

Derfor er $\mu_X = \mu_{S_1} + \cdots + \mu_{S_n} = np$ og siden $S_1, \cdots, S_n$ er uavhengige, $\sigma^2_X = \sigma^2_{S_1} + \cdots + \sigma^2_{S_n} = np(1-p)$.

b)

$$\begin{aligned} P(Z \leq 1) &= P(Z = 0) + P(Z = 1) \\ &= P(X = 0, Y = 0) \\ &= P(X = 1, Y = 0) + P(X = 0, Y = 1). \end{aligned}$$

Siden X og Y er uavhengige er dette

$$\begin{aligned} P(Z \leq 1) &= P(X = 0)P(Y = 0) \\ &= P(X = 1)P(Y = 0) + P(X = 0)P(Y = 1). \end{aligned}$$

Fra tabell P(X=0)=0.0168, P(X=1)=0.3602,P(Y=0)=0.0313,P(Y=1)=0.1563. derfor er $P(Z \leq 1) = 0.0168 \times 0.0313 + 0.3602 \times 0.0313 + 0.0168 \times 0.1563 = 0.0428$.

SLUTT