UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Onsdag 2. desember 2015
Tid for eksamen:	09.00 – 13.00
Oppgavesettet er på 3 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator, ordliste for STK1000 og lærebok (alle utgaver og det er lov til å notere i læreboken).

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

Oppgave 1

a

Forventningen $\mu_X = np = 1000 \star 0.02 = 20$ der $p = 0.02$ er sannsynligheten for tvillingfødsel og $n = 1000$ er antall fødsler. Dermed blir variansen til X gitt ved $\sigma_X^2 = np(1-p)$ og standardavviket $\sigma_X = \sqrt{np(1-p)} = \sqrt{20 \star 0.98} = 4.43$

b

Normaltilnærmingen for X er $\mathrm{N}(20, 4.43)$. Dermed får vi tilnærmet at sannsynligheten $\mathrm{P}(X \leq 11) \approx \mathrm{P}(Z \leq (11-20)/4.43) = \mathrm{P}(Z \leq -2.03) \approx 0.021$ når Z er en standardnormalfordelt tilfeldig variabel.

Siden både $np = 20 > 10$ og $n(1-p) = 980 > 10$ skal tilnærmelsen i teorien være god.

c

Med $q = 1 - \mathrm{P}(X \leq 11) = \mathrm{P}(X > 11)$ som sannsynligeten for flere enn 11 tvillingfødsler i et år har vi (siden årene er uavhengige) at sannsynlighet for at det alle 10 årene er mer enn 11 tvillingfødsler gis ved q^{10}. Med $p = 0.021$ gir dette $q^{10} \approx (1 - 0.021)^{10} = 0.808$.

Alternativt, fra Tabell C finner vi med $p = 0.02$ at sannsynligheten tilnærmet er lik 0.817.

Oppgave 2

a

Gjennomsnittet $\bar{D} = \frac{1}{10}\sum D_i$ har samme forventning μ som hver enkelt observasjon D_i og er derfor forventningsrett.

(Fortsettes på side 2.)

Samtidig har gjennomsnittet standardavvik $\sigma/\sqrt{n} = \sigma/\sqrt{10}$ der σ er standardavviket til D_i-ene.
Dessuten siden vi har antatt at D_i-ene er normalfordelte får vi eksakt at $\bar{D}$ er normalfordelt med forventning μ og standardavvik $\sigma/\sqrt{10}$.

b

Nullhypotesen blir $H_0 : \mu = 0$ og alternativet blir tosidig $H_0 : \mu \neq 0$.

Teststatikken som brukes er $t = \bar{D}/\text{SE}$ der $\text{SE} = s/\sqrt{10}$ er standardfeilen til $\bar{D}$, dvs. standardavviket innsatt estimatet s for σ. Her blir $t = 0.5/0.619 = 0.81$. Under nullhypotesen er denne statistikken t-fordelt med $n - 1 = 9$ frihetsgrader.

Den tilhørene p-verdien er oppgitt til 0.44 hvilket er godt over vanlige valg for signifikansnivå (som f.eks. er 0.05). Dermed kan vi ikke forkaste nullhypotesen og det er ikke grunnlag for å påstå at vi ikke kan stole på værmeldingene.

c

Intervallet beregnes ved formel $\bar{D} \pm t^\star SE$ der $t^\star = 2.26$ er 97.5-persentilen t-fordelingen med 9 frihetsgrader.

Betraktet som tilfeldige variable har intervallet en sannsynlighet på 0.95 for å inneholde forventningen μ.

Her ligger verdien $\mu = 0$ (dvs. nullhypotesen) godt innenfor intervallet. Dette samsvaret med at vi ikke forkastet nullhypotesen, noe vi ville gjort hvis på 5%-nivå hvis intervallet ikke inneholdt $\mu = 0$.

Oppgave 3

a

I enkel lineær regresjon har vi at korrelasjonskoeffisenten kvadrert er lik forklart andel av variasjon, som er oppgitt til R-sq=0.3175 i Minitab-utskriften. Siden vi også ser at estimert stigningskoeffisient $b_1 = 0.04899 > 0$ får vi at korrelasjonen $r = +\sqrt{0.3175} = 0.563$.

b

Med Y_i lik FEV_1 for barn nr. i kan modellen formuleres som

$$Y_i \text{ er uavhengige og normalfordelte } \text{N}(\beta_0 + \beta_1 x_i, \sigma)$$

der x_i er vekten for barn nr. i. Her er β_0, β_1 og σ parametre som må estimeres. Nullhypotesen om at det ikke er samsvar gis ved $H_0 : \beta_1 = 0$ og alternativet om at det er en slik sammenheng blir $H_a : \beta_1 \neq 0$. Vi finner en t-statistikk lik $t = b_1/SE_1 = 0,04899/0,00694 = 7.06$ som er t-fordelt med $n - 2 = 107$ frihetsgrader når $\beta_1 = 0$. I henhold til utskriften er p-verdien for testen mindre eller lik 0.0005, det er altså sterk grunn til forkaste nullhypotesen, dvs. å påstå at det er samvar mellom vekt og FEV_1.

(Fortsettes på side 3.)

c

Estimert forskjell blir lik b_1. Denne har en standardfeil $SE_1 = 0.00694$. For $n = 107$ frihetsgrader blir persentilene tilnærmet like (eller litt mindre enn) dem med 100 frihetsgrader. Fra Tabell D finner vi at 97.5 persentilen i t-fordelingen med 100 frihetsgrader er lik $t^\star = 1.984$. Dermed blir 95% konfidensintervallet

$$b_1 \pm t^\star SE_1 = (0.0352, 0.0627).$$

d

Vi ser at b_1, dvs. estimert koeffisienten for vekt, er klart mindre enn den var i den enkle lineære regresjonen. Dessuten blir samsvaret med vekt insignifikant, siden p-verdien for $H_0 : \beta_1 = 0$ er lik $0.071 > 0.05$ som er standard signifikansnivå.

(Samtidig ser vi at samsvaret med høyde er ganske stertk og faktisk har en t-verdi større enn den for vekt i den enkle lineære regresjonen.)

Hvis vi altså kjenner høyde er ikke vekt en særlig predikerende forklaringsvariabel. Vi kan si at i den enkle lineære regresjonen var høyde en *lurkende* variabel for vekt.

Fenomenet vi observerer skyldes at høyde og vekt er korrelerte variable, korrelasjonen er som vi ser så høy som 0.637.

SLUTT