UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK1110 — Statistiske metoder og dataanalyse 1
Eksamensdag:	Onsdag, 2. desember 2009
Tid for eksamen:	9.00 – 12.00
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell for χ^2-fordeling, tabell for t-fordeling
Tillatte hjelpemidler:	Godkjent lommeregner og Formelsamling for STK1100 og STK1110

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

Oppgave 1

En investor vil saksøke aksjemegler A fordi han mener avkastningen på porteføljen megleren forvalter for ham, er for lav. Vi har registrert månedlig avkastning (i prosent) for de 36 månedene megleren har tatt hånd om porteføljen. Vi ønsker å sammenligne med månedlig avkastning for de samme 36 månedene for en tilsvarende portefølje forvaltet av en annen megler B. Vi parrer observasjonene (diff = megler A - megler B) hver måned, og antar at differansene i avkastning for de 36 månedene kan betraktes som uavhengige av hverandre. Gjennomsnittet av de 36 månedlige avkastningene for megler A var -1.10 prosent, tilsvarende for megler B var 0.95 prosent.

Empirisk standardavvik for de 36 differansene var 5.89 prosent. Et normalfordelingsplott av de 36 differansene viser ingen ekstreme verdier eller skjevheter i fordelingen.

a) Anta at fordelingen til differansene har forventning μ og standardavvik σ. Forklar hvorfor det er naturlig å teste hypotesene

$$H_0 : \mu = 0 \quad \text{mot} \quad H_a : \mu < 0.$$

Finn en passende testobservator, og spesifiser dennes fordeling når nullhypotesen er sann. Gjør rede for hvilke antakelser du legger til grunn.

b) Beregn forkastningsområdet svarende til at testen skal ha nivå 0.01, og

(Fortsettes på side 2.)

konkluder på bakgrunn av tallene i oppgaveteksten. Beregn også tilhørende P-verdi (så godt du kan fra vedlagte tabell). Bør investoren gå til søksmål?

c) Beregn et 95% konfidensintervall for standardavviket σ.

d) Det ble også vurdert å bruke en to-utvalgstest istedenfor testen basert på de parrede observasjonene. Hvorfor ville to-utvalgstesten være et bedre alternativ under visse forutsetninger, og hvorfor tror du den parrede testen ble brukt i dette tilfellet?

Oppgave 2

En ny antikoagulant (legemiddel mot blodpropp) er under utvikling. Som et ledd i utprøvingen fikk 12 friske menn og 8 friske kvinner antikoagulanten i ulike doser (i milligram), og prothrombin-tid ble målt for hver av dem (i sekunder). Prothrombin-tid er et mål som sier noe om hvor raskt blodet koagulerer. Man ønsker å bruke datamaterialet til å belyse sammenhengen mellom dose av antikoagulanten, kjønn og prothrombin-tid ved å utføre en regresjonsanalyse med dose x_1 og kjønn x_2 som forklaringsvariable og prothrombin-tid som respons Y.

I figuren på neste side er prothrombin-tid plottet mot dose, med forskjellige symboler for kvinner og menn. Regresjonsmodellen vi vil bruke er

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{1}$$

der x_{i1} er dose for person i og $x_{i2} = 0$ hvis person i er en mann og $x_{i2} = 1$ hvis person i er en kvinne. ϵ_i er uavhengige og normalfordelte $N(0, \sigma^2)$, $i = 1, \cdots, 20$.

a) Gi en tolkning av parametrene β_1 og β_2. Virker modellen rimelig i forhold til plottet av dataene? Begrunn svaret.

Vi skal uansett gå videre med modell (1) i punkt b) og c). En analyse av datasettet gir følgende estimater ved minste kvadraters metode: $\hat{\beta}_0 = 7.155$, $\hat{\beta}_1 = 0.05849$ og $\hat{\beta}_2 = 3.7866$. Estimert standardfeil er beregnet til hhv. $s_{\hat{\beta}_0} = 2.445$, $s_{\hat{\beta}_1} = 0.01201$ og $s_{\hat{\beta}_2} = 0.3886$.

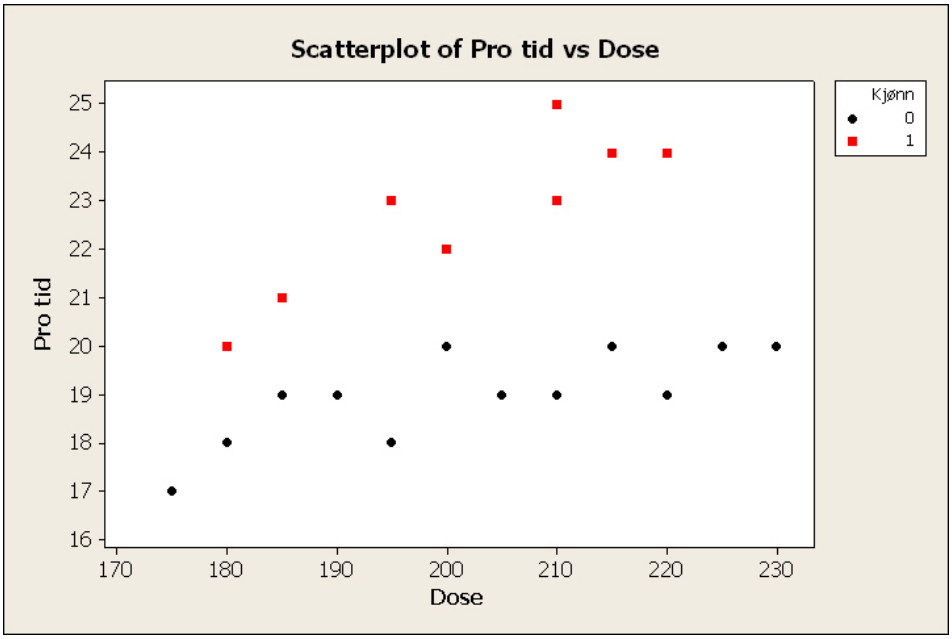
b) Beregn et 95% konfidensintervall for forventet forskjell i prothrombin-tid mellom kvinner og menn ved lik dose. Bruk intervallet til å teste om denne forskjellen er signifikant forskjellig fra 0. Hva blir signifikansnivået for denne testen?

c) Hvordan beregnes residualene i denne modellen (du skal ikke beregne dem!)? Beskriv kort forskjellige plott som kaster lys over hvor godt modellen

(Fortsettes på side 3.)

passer til data. Multippel R^2 er beregnet til 87.3 %. Hvordan skal dette tallet tolkes?

d) Finn et uttrykk for endringen i forventet prothrombin-tid når dosen økes med d milligram. Vis at denne endringen er den samme for kvinner og menn. Hvordan vil du modifisere regresjonsmodellen dersom forventet endring i prothrombin-tid for en gitt endring i dose antas forskjellig for kvinner og menn? Se igjen på figuren og diskuter kort om dette er en rimelig utvidelse av modellen.



Oppgave 3

I frykt for en pandemisk influensa planlegger Folkehelseinstituttet å massevaksinere befolkningen. Det er ønskelig at mer enn 60% av befolkningen vaksineres. På et gitt tidspunkt utføres en spørreundersøkelse for å finne ut hvor stor vaksinasjonsviljen er i befolkningen. Anta at p er andelen av befolkningen som sier ja til vaksinen. Et tilfeldig utvalg på n personer trekkes ut, og disse må svare på om de vil vaksineres eller ikke. La X være den tilfeldige variabelen som beskriver antallet som sier ja blant de n. Vi er interessert i å estimere p.

a) Vis at den intuitive estimatoren for p, $\hat{p} = X/n$, faktisk er sannsynlighetsmaksimerings-estimatoren (maximum likelihood estimator) for p. Er den også en momentestimator for p?

b) Finn forventning $E(\hat{p})$ og varians $V(\hat{p})$ for $\hat{p}$ og vis at estimatoren er

(Fortsettes på side 4.)

konsistent, altså at $\hat{p}$ konvergerer i sannsynlighet mot p når n vokser. (Tips: Bruk Chebyshevs ulikhet fra formelsamlingen.)

c) Anta at utvalgsstørrelsen n er så stor at fordelingen til $\hat{p}$ kan tilnærmes med en normalfordeling. Utled og gjennomfør en hypotesetest for å undersøke om innsamlede data viser at p er så stor som Folkehelseinstituttet ønsker. Du kan bruke at $n = 200$ og at 122 av disse svarte ja.

d) Finn sannsynlighetsmaksimerings-estimatorer for $E(X)$ og $V(X)$. Vis at den resulterende estimatoren for $V(X)$ ikke er forventningsrett, og foreslå en modifisert estimator som er det.

SLUTT