Andre sett med obligatoriske oppgaver i STK1110 høsten 2013

Dette er andre sett med obligatoriske oppgaver i STK1110 høsten 2013. Oppgavesettet består av fire oppgaver. I oppgave 2 og 3 skal du gjøre beregningene 'for hånd', mens i oppgave 1 og 4 skal du også bruke R. Gjør oppgave 4 til sist, da det er helt nytt stoff fra kapittel 12. Oppgave 4 a) og b) er del av det obligatoriske settet, mens 4 c) og d) er frivillige. Der du bruker R (eller et annet program), må utskrifter legges ved/limes inn. Prøv å besvare spørsmålene kort og konsist, men likevel med gode forklaringer!

Du må bruke Matematisk institutts forside ved innlevering. Det er helt i orden og utmerket dersom dere samarbeider og diskuterer hvordan oppgavene skal løses, men utformingen og formuleringen av besvarelsene må være individuelle. Hvis flere studenter samarbeider om å løse oppgavene, må hver student levere sin *selvstendige* besvarelse, og det må gå frem av besvarelsen hvem den enkelte har samarbeidet med. Se ellers "Regelverk for obligatoriske oppgaver" som er gitt på kursets hjemmeside.

Besvarelsen leveres ved ekspedisjonen til Matematisk institutt, 7. etasje, Niels Henrik Abels hus.

Frist for innlevering er torsdag 7. november kl. 14.30.

Oppgave 1

Følgende tabell viser 9 målinger av kroppstemperatur for friske kvinner, $x_1, \ldots, x_9$, og 9 målinger for friske menn, $y_1, \ldots, y_9$. Hensikten med denne oppgaven er å undersøke om det er tilstrekkelig informasjon i tabellen til å kunne konkludere med at kroppstemperaturen er forskjellig for friske menn og kvinner.

Kroppstemperatur	
Menn	Kvinner
36.1	36.6
36.3	36.7
36.4	36.8
36.6	36.8
36.6	36.8
36.6	37.0
36.7	37.1
37.0	37.3
37.1	37.4

a) Lag boksplott som viser fordelingen av observasjonene. Kommenter hva du finner.

b) Lag normalfordelingsplott for de to observasjonssettene. Kommenter hva du ser.

I resten av oppgaven antar vi at observasjonene er realisasjoner av normalfordelte variable. I c) og d) skal du forklare hvordan tester og konfidensintervaller konstrueres, og sette inn i formlene du utleder. Sjekk deretter svarene du får mot R-prosedyren t.test().

c) Anta at variansen er den samme for de to utvalgene, og test med nivå 5% om det er noen forskjell i forventet kroppstemperatur. Beregn p-verdien, og lag et 95% konfidensintervall for forventet forskjell.

d) Gjennomfør testen og beregn p-verdien også i det tilfellet der man ikke antar felles varians. Diskuter og forklar resultatene.

e) Utled og gjennomfør en F-test for å sjekke om det er noen grunn til å påstå at variansene er forskjellige. Sjekk mot var.test() i R.

f) Se nå på situasjonen der man vurderer å innhente to nye målinger. La X_{10} være verdien for kvinne, Y_{10} for mannen, slik at forskjellen er $X_{10} - Y_{10}$. Vi antar nå at alle observasjonene er normalfordelte med samme varians. Begrunn at et rimelig anslag for $X_{10} - Y_{10}$ er differansen mellom gjennomsnittet av de 9 eksisterende målingene for kvinner og menn, $\bar{X} - \bar{Y}$.

   Hva er fordelingen til $X_{10} - Y_{10} - (\bar{X} - \bar{Y})$? Bruk dette til å lage et intervall, som er slik at sannsynligheten er 0.95 for at $X_{10} - Y_{10}$ vil ligge i intervallet. Slike intervaller kalles *prediksjonsintervaller* (gjennomgått for ett-utvalgssituasjon på forelesning).

   Forklar hva forskjellen er mellom et slikt intervall og et konfidensintervall for $\mu_X - \mu_Y$. Hvordan skal et prediksjonsintervall tolkes?

   [Hint: Siden alle variablene er normalfordelte, er også $X_{10} - Y_{10} - (\bar{X} - \bar{Y})$ det. Det er derfor nok å beregne forventning og varians for å finne fordelingen til denne størrelsen.]

Oppgave 2

Siden eneggede tvillinger har samme genetiske materiale, brukes såkalte tvillingstudier til å kartlegge hvordan miljøet virker inn på ulike egenskaper. I en bok av den amerikanske forskeren Susan Faber finner vi data for $n = 31$ tvillingpar

der den ene tvillingen vokste opp hos biologiske foreldre (Twin A) og den andre vokste opp hos andre familiemedlemmer, foster- eller adoptiv-foreldre (Twin B). Nedenfor finnes en oppsummering av målt IQ for disse personene. Spørsmålet vi ønsker å belyse er om det er forskjell i IQ hos eneggede tvillinger der den ene tvillingen har vokst opp hos biologiske foreldre, og den andre ikke.

	N	Mean	StDev	SE Mean
Twin A	31	93.32	15.41	2.77
Twin B	31	96.58	13.84	2.49
Difference	31	-3.26	8.81	1.58

a) Begrunn hvorfor en paret sammenligning er best egnet i denne situasjonen. Beskriv kort hvilke antakelser vi må legge til grunn for videre analyse.

b) Kall forventet forskjell mellom Twin A og Twin B for μ_D. Sett opp nullhypotese og alternativ hypotese for å besvare spørsmålet om forskjell i IQ. Finn en egnet testobservator, og beregn dennes numeriske verdi. Beregn så tilhørende p-verdi. Spesifiser antall frihetsgrader i fordelingen du bruker. Formuler din konklusjon på testen.

c) Finn et 95% konfidensintervall for μ_D. Hva betyr det at dette intervallet dekker kun negative verdier? Forklar kort om sammenhengen mellom tosidig testing og konfidensintervaller.

Oppgave 3

En undersøkelse presentert i Aftenposten slår opp på førstesiden at småbarnsfedrene nå opplever tidsklemmen (mellom familie og arbeidsliv) sterkere enn småbarnsmødrene. Undersøkelsen bygger på intervju med 3000 kvinner og 3000 menn som har barn i rett alder. 16.2 % av fedrene (dvs. 486 personer) opplever ofte tidsklemmeproblemer, mens 14.7% (dvs. 441 personer) av mødrene opplever det samme. Er forskjellen mellom mødre og fedre signifikant? Formuler hypoteser og beregn en p-verdi, og konkluder. Kommenter kort.

Oppgave 4

Tabellen nedenfor angir 18 målinger av snømengde om vinteren i et fjellområde og vannstanden i en elv i samme område etter snøsmelting om våren. De 18 målingene representerer 18 sesonger spredd over en lengre tidsperiode. Her er vannstanden, som er angitt i tommer, respons- eller avhengig variabel, mens snømengden, målt ved noe som heter vannekvivalens, er forklaringsvariabel. Sammenhengen mellom snømengde og vannstand er viktig for bl.a. prediksjon av

vannføring og flomfare. Dataene finnes i filen 'dataoblig2.txt' på kurshjemmesiden.

Snømengde	Vannstand
23.1	10.5
32.8	16.7
31.8	18.2
32.0	17.0
30.4	16.3
24.0	10.5
39.5	23.1
24.2	12.4
52.5	24.9
37.9	22.8
30.5	14.1
25.1	12.9
12.4	8.8
35.1	17.4
31.5	14.9
21.1	10.5
27.6	10.5
27.6	16.1

a) Beskriv en enkel lineær regresjonsmodell for sammenhengen mellom snømengde og vannstand. Tilpass en regresjonslinje til dataene ovenfor ved hjelp av R-funksjonen lm(). Plott observasjonene og den tilpassede regresjonslinja. Virker estimatene for koeffesientene rimelige?

b) Plott residualene mot forklaringsvariabelen. Lag også et normalfordelingsplott for residualene. Hvordan vurderer du modellens egnethet?

c) Beregn et estimat for variansen til feilleddene. Konstruer et 95% konfidensintervall for stigningstallet β_1.

d) Utled en test for $H_0 : \beta_0 = 0$ mot $H_1 : \beta_0 \neq 0$. Bruk nivå 5%. Du kan hente formelen for variansen til $\hat{\beta}_0$ fra formelsamlingen. Gjennomfør testen. Hva er p-verdien? Sammenlign med resultat fra lm().