UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 – Innføring i anvendt statistikk
Eksamensdag:	Tirsdag 29. november 2016
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 4 sider	
Tillatte hjelpemidler:	Godkjent kalkulator, ordliste for STK1000, og lærebok (alle utgaver, og det er lov å notere i læreboka)

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.
Alle deloppgaver teller likt i vurderingen av besvarelsen. Lykke til!

Oppgave 1

I et tilfeldig utvalg på 1000 normalvektige personer, og 1000 overvektige personer, måles konsentrasjonen av 200 ulike proteiner i blodet. For omtrent halvparten av disse proteinene ser verdiene ut til å være relativt normalfordelte (i begge gruppene). For resten av proteinene ser verdiene ut til å være skjevfordelte (i begge gruppene), og for 14 av proteinene er fordelingen ekstremt skjev. Verdiene til de 14 sistnevnte proteinene er knapt målbare for majoriteten av de 2000 personene i studien, mens et fåtall personer i hver gruppe har ekstremt høye verdier.

a) Hvilke oppsummeringstall (deskriptiv statistikk) ville du brukt for å beskrive konsentrasjonen av disse proteinene i blod? Begrunn svaret.
b) Sett opp nullhypotese og alternativ hypotese for å undersøke om det er forskjell i konsentrasjonen av proteiner i blodet til deltakerne i de to gruppene. Hvilke(n) hypotesetest(er) vil du bruke for å sammenligne gruppene? Begrunn svaret.
c) Forskerne som planla denne studien ønsket å bruke et signifikansnivå på 5%, altså $\alpha=0.05$. Hva betyr dette?
   Anta at H_0 er sann for alle de 200 proteinene. Hvor mange signifikante gruppeforskjeller kan man allikevel forvente å finne? (Dersom man gjør 200 uavhengige hypotesetester, hver med signifikansnivå 5%?)
   Anta så at man gjør en studie der man ikke gjør tester for alle de 200 proteinene, men

velger tre av dem, som antas å være uavhengige. Hvis signifikansnivået er 5% i hver test, hva er den totale sannsynligheten for type 1-feil i denne studien?

Oppgave 2

I en test av meterstokker/tommestokker som Forbrukerrådet gjorde i 2016, ble 21 meterstokker vurdert etter hvor nøyaktige de var. Med hjelp fra Justervesenet ble meterstokkene festet i en kalibrert rigg, og så ble punktet der meterstokken viste 99 cm sammenlignet med fasit ved hjelp av laserinferometer. Forbrukerrådet ønsket å teste om nøyaktigheten på meterstokkene hadde en sammenheng med prisen.

De to utskriftene under viser en korrelasjonsanalyse og en regresjonsanalyse for sammenhengen mellom pris (i kr), y, og nøyaktighet (i mm), x.

```
> cor.test(pris,avvik_mm)

        Pearson's product-moment correlation

data:  pris and avvik_mm
t = 0.86669, df = 19, p-value = 0.3969
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2584288  0.5780384
sample estimates:
      cor
0.1950137

> summary(lm(avvik_mm~pris))

Call:
lm(formula = avvik_mm ~ pris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97907 -0.26104 -0.01009  0.24867  0.82619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.160012   0.152956   1.046    0.309
pris        0.001752   0.002021   0.867    0.397

Residual standard error: 0.423 on 19 degrees of freedom
Multiple R-squared:  0.03803,  Adjusted R-squared:  -0.0126
F-statistic: 0.7511 on 1 and 19 DF,  p-value: 0.3969
```

a) Formulér regresjonsmodellen som er utgangspunktet for regresjonsanalysen. Ta så utgangspunkt i de to analysene i utskriftene over, og formulér de tilhørende to sett med hypoteser (altså nullhypotese og alternativ hypotese), for to ulike parametere, som begge kan brukes når man vil teste om det er en sammenheng mellom pris og nøyaktighet. Begrunn hvorfor du velger ensidig eller tosidige hypoteser. Hvilke(n) konklusjon(er) trekker du?

Oppgave 3

I en studie av 200 friske gravide kvinner ble deltakerne rekruttert etter hvert som de søkte fødeplass på et gitt sykehus. Forskerne ønsket å finne ut om det var en sammenheng mellom

mors blodsukkernivå (målt i mmol/l) og barnets fødselsvekt (målt i gram). På inklusjonstidspunktet var kvinnene gravide i tredje måned, og fastende blodsukker ble målt. Det måles om morgenen før frokost. Følgende regresjonsanalyse ble gjort:

```
> summary(lm(fodselsvekt~blodsukker))

Call:
lm(formula = fodselsvekt ~ blodsukker)

Residuals:
    Min      1Q  Median      3Q     Max
-895.23 -366.63  -46.74  393.25 1572.35

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3149.89     331.43   9.504   <2e-16 ***
blodsukker    171.69      81.93   2.096   0.0374 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 488.9 on 198 degrees of freedom
Multiple R-squared:  0.0217,   Adjusted R-squared:  0.01676
F-statistic: 4.391 on 1 and 198 DF,  p-value: 0.0374
```

a) Hva er effektmålet her, og hvordan tolkes det? Gi et estimat for sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt, og beregn et 95% konfidensintervall for det samme.

Bakgrunnen for studien var at man allerede visste at gravide kvinner med diabetes føder tyngre barn enn gravide uten diabetes. Også kvinner med svangerskapsdiabetes (den type diabetes som oppstår i løpet av et svangerskap, oftere blant overvektige kvinner, men som forsvinner etter fødselen) føder også større barn enn gjennomsnittet. Man vet også at kvinner med svangerskapsdiabetes har økt risiko for type 2- diabetes senere i livet. Overvekt, både blant menn og kvinner, er også forbundet med økt risiko for type 2-diabetes, og man antar at noen av mekanismene involvert i dette er betennelsesreaksjoner i kroppen som oppstår ved overvekt, og at disse påvirker kroppens evne til å regulere blodsukkeret.

b) Hva menes med en konfunderende variabel (confounder eller lurking variable)? Kan mors body mass index (bmi), altså (vekt i kg)/(høyde i m)^2, sies å være en konfunderende variabel for sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt? Begrunn svaret.
c) Bruk følgende utskrift til å gi et nytt estimat og et nytt 95% konfidensintervall for sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt. Er det en sammenheng mellom mors blodsukkernivå og barnets fødselsvekt? Begrunn svaret.

```
> summary(lm(fodselsvekt~blodsukker+bmi))

Call:
lm(formula = fodselsvekt ~ blodsukker + bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-922.91 -401.80   -7.42  408.93 1566.63
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2964.105    336.907   8.798 7.03e-16 ***
blodsukker    93.763     87.426   1.072   0.2848
bmi           19.883      8.397   2.368   0.0189 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 483.3 on 197 degrees of freedom
Multiple R-squared:  0.04877,      Adjusted R-squared:  0.03911
F-statistic:  5.05 on 2 and 197 DF,  p-value: 0.007262
```

Oppgave 4

I en studie av øretermometere fant man ut at sammenhengen mellom den sanne kroppstemperaturen (sentraltemperaturen) y_i, og målingene fra øretermometeret x_i (kalt `ear` i utskriften), kunne uttrykkes ved regresjonsligningen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma)$$

Gjennomsnitt og standardavvik (i ºC) var $\bar{x} = 37.11$, $sd_x = 0.83$, og $\bar{y} = 37.89$, $sd_y = 0.92$.

Utskriften viser en regresjonsanalyse som ble gjort på målinger av 237 intensivpasienter, der det var mulig å gjøre en nøyaktig måling av sentraltemperaturen:

```
> summary(lm(central~ear))

Call:
lm(formula = central ~ ear)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25600 -0.31566 -0.05978  0.28854  1.60031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.74544    1.51257   2.476    0.014 *
ear          0.92017    0.04075  22.580   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5172 on 235 degrees of freedom
Multiple R-squared:  0.6845,   Adjusted R-squared:  0.6832
F-statistic: 509.9 on 1 and 235 DF,  p-value: < 2.2e-16
```

a) Gi estimater og tolkning av estimatene for parameterne β_0 og β_1, og sett opp hypotesene i de hypotesetestene som reflekteres i de to første p-verdiene i utskriften.
b) Lag et 95% prediksjonsintervall for sentraltemperaturen når øretemperaturen viser 38 ºC.
c) Differansene mellom målingene av sentraltemperaturen og øretemperaturen hadde et gjennomsnitt på 0.78 ºC og et standardavvik på 0.52 ºC. Beregn et 95% konfidensintervall for forventet forskjell på de to målemetodene, og kommentér svaret.

SLUTT