EKSAMENSOPPGAVE

Institutt: **IKBM**

Eksamen i: **STAT 100** **STATISTIKK**

Tid: Fredag 16.mai 9:00 – 12:30 (3.5 timer)

Emneansvarlig: Solve Sæbø, 6496 5823

Tillatte hjelpemidler:
C3: alle typer kalkulatorer, alle andre hjelpemidler

Oppgaveteksten er på: 15 (inkludert F-tabell og svarskjema)

antall sider inkl. vedlegg

Alle deloppgaver teller likt. For hver oppgave er det 5 svaralternativer. Kun ett svar-alternativ er riktig. Du får ett poeng for riktig svar, null poeng for feil svar. Maksimal score er da 40 poeng.

Alle svar føres i svarskjemaet på side 12, og denne siden er den ENESTE som skal leveres når eksamen er slutt. Husk å skrive kandidatnummer på svarskjemaet!

Kvalitet på drikkevann (Oppgave 1-7)
En undersøkelse av drikkevannskvalitet ble utført i en landsby i Afrika. Dagens drikkevann blir hentet fra lokalitet A i en innsjø, men det vurderes å flytte inntaket til lokalitet B i den samme sjøen hvis kvaliteten der er bedre. Et lokalt miljøbevisst ektepar fikk i oppgave å ta månedlige prøver i en periode på 10 måneder fra begge lokaliteter og måle mengden av bakterien E.coli. Konsentrasjonen av E.coli-bakterien i målingene ble:

Mnd :	Jan	Feb	Mar	Apr	Mai	Jun	Jul	Aug	Sep	Okt
Lok. A:	212	210	217	205	212	201	238	225	219	198
Lok. B:	202	198	190	195	205	216	220	214	206	188

La X være den tilfeldige variabelen for E.coli-konsentrasjonen for lokalitet A med fordelings-antagelse $X \sim N(\mu_A, \sigma_A)$ og tilsvarende la Y være E coli-konsentrasjonen for lokalitet B med fordelingsantagelse $Y \sim N(\mu_B, \sigma_B)$. Man ønsker å bruke målingene til å teste om forventet mengde E.coli-bakterier ved lokalitet B er mindre enn ved lokalitet A.

Oppgave 1
I henhold til teksten ovenfor, hvilke hypoteser ønsker man å teste?

a) $H_0 : \mu_A = \mu_B$ mot $H_1 : \mu_A \neq \mu_B$
b) $H_0 : \mu_A - \mu_B = 0$ mot $H_1 : \mu_A - \mu_B < 0$
c) $H_0 : \sigma_A = \sigma_B$ mot $H_1 : \sigma_A > \sigma_B$
d) $H_0 : \mu_A = \mu_B$ mot $H_1 : \mu_A < \mu_B$
e) $H_0 : \mu_A - \mu_B = 0$ mot $H_1 : \mu_A - \mu_B > 0$

Oppgave 2
Her er det lurt å betrakte dataene som parvise data i en statistisk analyse. Hva er den mest korrekte begrunnelsen for dette?

a) Fordi vi har målt konsentrasjonen på to ulike lokaliteter, A og B.
b) Fordi et ektepar har utført målingene.
c) Fordi målinger gjort for en gitt måned antageligvis er ganske like de målinger som ble gjort måneden før.
d) Fordi det vil kunne fjerne forstyrrende måned-til-måned variasjon i E.coli målingene før man tester hypotesene.
e) Fordi det vil kunne fjerne forstyrrende variasjon mellom lokalitetene før man tester hypotesene.

Oppgave 3
Man ville analysere dataene som parvise og beregnet nye observasjoner som $D_i = X_i - Y_i$ for alle månedsnumre $i = 1, \ldots, 10$, og man antok modellen $D \sim N(\mu_D, \sigma_D)$ og uavhengighet mellom alle D_i. De forventningsrette estimatene for modellparametrene basert på data var $\hat{\mu}_D = 10.30$ og $\hat{\sigma}_D = 10.54$. Hva blir verdien på den t-fordelte testobservatoren som skal brukes for å teste om lokalitet B har lavere forventet E.coli-konsentrasjon enn lokalitet A?

a) 0.98 b) 3.09 c) 3.45 d) 2.55 e) 1.15

Oppgave 4
Hvor mange frihetsgrader har testobservatoren i oppgave 3?

a) 9 b) 2 c) 8 d) 10 e) 18

Oppgave 5
Dersom de to lokalitetene har samme forventede nivå av E.coli-bakterier bør sannsynligheten for at B har lavere måling enn A være lik *p*=0.5. En annen måte å sammenlikne de to lokalitetene er derfor å telle opp hvor ofte i løpet av *n*=10 måneder at lokalitet B har lavere nivå enn lokalitet A og teste om man kan forkaste en $H_0 : p = 0.5$ og heller hevde $H_1 : p > 0.5$. Dersom alternativ hypotese er riktig bør man altså skifte drikkevannskilde. Fra data ser vi at lokalitet B er bedre enn A i 9 av 10 måneder. Hva blir den *eksakte* p-verdien for hypotesetesten?

a) 0.0099 b) 0.001 c) 0.011 d) 0.989 e) 0.022

Oppgave 6
Anta at noen (feilaktig) besluttet å analysere E.coli målingene fra lokalitet A og B som *to uavhengige utvalg*. De ville fortsatt teste om lokalitet B har lavere forventet nivå enn lokalitet A. Hva blir absoluttverdien (tallverdien) på den t-fordelte testobservatoren i dette tilfellet? (Du kan bruke følgende utskrift fra R Commander som hjelp i utregningene)

```
          mean       sd         var     n
lok A     213.7      11.83      140.01  10
lok B     203.4      10.93      119.38  10
```

a) 1.97 b) 2.45 c) 2.02 d) 3.04 e) 0.89

Oppgave 7
Hvor mange frihetsgrader har testobservatoren i oppgave 6?
a) 19 b) 20 c) 8 d) 10 e) 18

Gjødsling av tomatplanter (Oppgave 8-16)
Man ønsker å sammenlikne *k*=3 typer gjødsel (Type A, B og C) for tomatplanter. Et forsøksfelt for tomat-dyrking ble delt inn i tre deler, og ved loddtrekning ble det bestemt hvilken gjødseltype som skulle brukes på hver av delene. Dette ble gjentatt i fire år (sesonger), og man målte total avling (*Y*) av tomat for hver gjødseltype alle årene. Vi vil anta at resultatene fra ulike år er uavhengige. Følgende modell antas for avlingsdataene: $Y_{ij} = \mu_i + \epsilon_{ij}$, der $\epsilon_{ij} \sim N(0, \sigma)$

Videre er μ_i for $i = 1, 2, 3$ forventningen for hhv gjødseltype A, B og C, og $j = 1, 2, 3, 4$ angir observasjonsår.
Deler av resultatene av en statistisk analyse av dataene er gitt i følgende R Commander utskrift (noen tall er erstattet med *):

```
> Anova(LinearModel)
          Df Sum Sq Mean Sq F value  Pr(>F)
Type       *      *  7.4533       * 0.02098
Residuals  9 10.960  1.2178

     mean       sd        n
A    25.0       1.058      4
B    24.7       1.160      4
C    22.5       1.089      4
```

Oppgave 8
Hvor stor er SS_G (dvs kvadratsummen som beskriver variasjon mellom gjødseltypene)?
a) 12.224 b) 14.907 c) 3.727 d) 22.359 e) 17.986

Oppgave 9
Hvor stor er den F-fordelte testobservatoren for å teste hypotesen om at alle gjødseltypene har samme forventede avling?

a) 9.08 b) 4.26 c) 5.78 d) 0.16 e) 6.12

Oppgave 10
En av observasjonene med gjødseltype A hadde en avling på 26.0. Hva er residualet for denne observasjonen?

a) 0 b) -3.5 c) 1.06 d) 1.0 e) -0.3

Oppgave 11
Resultatene tilsier (med et testnivå på 5%) at minst to av gjødseltypene gir forskjellig forventning for tomatavlingen. Gjødseltypene A og B er begge kunstgjødsler, mens type C er naturlig. For å undersøke om kunstgjødsel generelt gir høyere avling enn naturgjødsel definerte man kontrasten

$$\theta = \frac{\mu_1 + \mu_2}{2} - \mu_3$$

Hva er et forventningsrett estimat for θ basert på de observerte data?

a) Ukjent siden μ'ene er ukjente b) 3.16 c) 2.05 d) 14.7 e) 2.35

Oppgave 12
Hvilket sett av hypoteser er korrekte for å utføre testen beskrevet i oppgave 11?

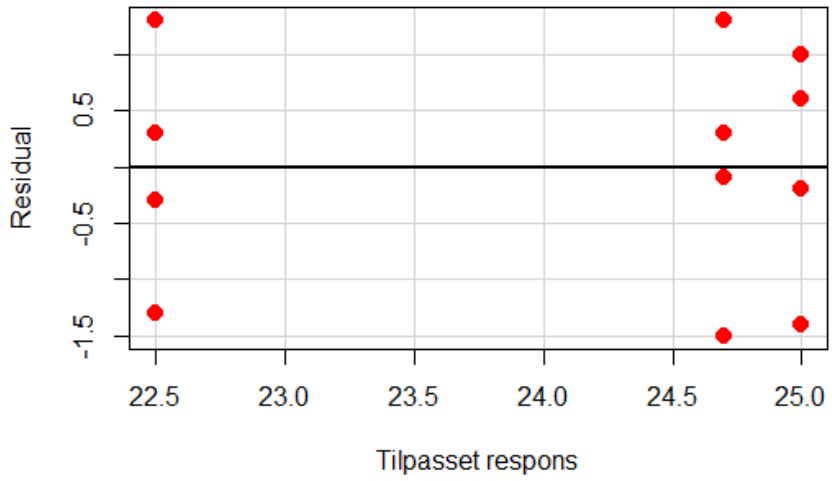
a) $H_0 : \theta = 0$ mot $H_1 : \theta \neq 0$
b) $H_0 : \theta \neq 0$ mot $H_1 : \theta = 0$
c) $H_0 : \theta = 0$ mot $H_1 : \theta < 0$
d) $H_0 : \theta = 0$ mot $H_1 : \theta > 0$
e) $H_0 : \theta \neq 0$ mot $H_1 : \theta < 0$

Oppgave 13
Man ville også undersøke om det var forskjell på kunstgjødseltypene mht forventet avling og definerte en annen kontrast ved $\theta = \mu_1 - \mu_2$. Hva blir standardfeilen til den forventningsrette estimatoren til denne kontrasten?

a) 0.78 b) 1.03 c) 0.53 d) 0.74 e) 0.92

Oppgave 14
Gitt at det testes på 5% nivå, hvorfor er det unødvendig å utføre en kontrastanalyse for å teste om det er forskjell på gjødseltype A og C?

a) Fordi disse to typene har nesten identisk standardavvik på hhv 1.058 og 1.089 og dermed sannsynligvis ikke vil være signifikant forskjellige fra hverandre.
b) Unødvendig å teste disse to fordi den ene er kunstgjødsel og den andre er naturgjødsel, og det blir som å sammenlikne epler og pærer.
c) Hvis to forventninger er ulike, som p-verdien tilsier, så må det i alle fall være mellom disse to siden de har størst forskjell mellom gjennomsnittene.
d) Fordi verdien 0 ikke ligger i intervallet [22.5, 25.0] må det garantert være en forskjell mellom de to gjødseltypene.
e) Fordi antall observasjoner er mindre enn 5.



Figur 1: Residual-plott for oppgave 15

Oppgave 15
Figur 1 viser residualene plottet mot de tilpassede responsverdiene fra den estimerte modellen for tomatavlingsdataene. Hvilken påstand knyttet til figuren er mest korrekt?

a) Figuren gir en klar indikasjon på at antagelsen om normalfordelte støyledd er feil.
b) Figuren gir en klar indikasjon på at antagelsen om likt støystandardavvik σ for alle gjødseltyper er feil.
c) At punktene ligger i tre klare vertikale bånd og ikke tilfeldig spredt er en klar indikasjon på at antagelsen om lineær modell er feil.
d) Figuren viser at det ikke er klare brudd på modellantagelsene om lineæritet og konstant varians.
e) Figuren viser at det er brudd på antagelsen om uavhengige støyledd siden punktene ligger i tre vertikale grupper.

Oppgave 16
Hvordan tolkes parameteren σ i modellen?

a) Den er et mål på variasjon i tomatavling innenfor en gitt type gjødsel.
b) Den er et mål på total variasjon i tomatavling sett over alle gjødseltyper.
c) Den er et mål på andelen av forklart variasjon i den estimerte modellen.
d) Den er et mål på variasjon i gjødseltype.
e) Den er et mål på forventet tomatavling for en gitt gjødseltype.

Mer om bakterier i vann (Oppgave 17-23)
Vann brukes til å rengjøre melkeutstyr på gårder, og vannkvaliteten ble undersøkt på k=4 gårder. Konsentrasjonen (Y) av bakterier ble målt i 5 prøver fra hver gård. Følgende modell ble antatt for konsentrasjonen for måling j (der j=1,…,5) på gård nummer i (der i=1,…,4): $Y_{ij} = \mu_i + \epsilon_{ij}$ hvor $\epsilon_{ij} \sim N(0, \sigma)$ og uavhengige.
Bruk R utskriften nedenfor fra en analyse av dataene i den grad du finner det nødvendig.

```
> Anova(LinearModel.3, type="II")
          Df Sum Sq Mean Sq F value   Pr(>F)
Gård       3   4820  1606.7  5.3556 0.009559
Residuals 16   4800   300.0

      mean       sd        n
Gård1  180    15.81        5
Gård2  190    18.71        5
Gård3  196        *        5
Gård4  222    19.24        5

Contrast:
                    Estimate Std. Error   t value
Gård c=( 0 1 0 -1 )      -32   10.95445 -2.921187
```

Oppgave 17

Hvis du vil teste om det er signifikant forskjell på minst to av gårdene mht forventet bakteriekonsentrasjon med testnivå $\alpha = 0.05$, hva er da den kritiske F_α-verdien som angir nedre grense for forkastningsområdet?

a) 8.70 b) 3.24 c) 3.01 d) 2.87 e) 3.10

Oppgave 18

Hva er IKKE riktig å si om testnivået α i en hypotesetest om at alle μ_i'ene i modellen er like?

a) Den er også kjent som sannsynligheten for «type 1 feil».
b) Det er sannsynligheten for feilaktig forkastning av nullhypotesen.
c) Det er sannsynligheten under H_0 for å observere en verdi av testobservatoren i forkastningsområdet.
d) Det er sannsynligheten under H_0 for at den alternative hypotesen aksepteres.
e) Det er sannsynligheten under H_0 for at den alternative hypotesen er korrekt.

Oppgave 19

Resultatene gir en forkastning av nullhypotesen om at alle forventninger er like. Man besluttet å sammenlikne forventninger parvis for å finne hvor forskjellene lå. For eksempel sammenliknet man forventet bakterienivå på gård 2 med gård 4 ved kontrasten $\theta = \mu_2 - \mu_4$. Som man ser nederst i utskriften så er $\hat{\theta} = -32$ og $SE(\hat{\theta}) = 10.95$. Hva blir et 99% konfidensintervall for den sanne θ?

a) [-71.11, 7.11]
b) [-60.56, -3.44]
c) [-63.98, -0.015]
d) [-49.23, -14.77]
e) [-42.95, -21.05]

Oppgave 20
Vi blir gitt samme opplysninger som i oppgave 19, men ønsker å teste på 5% nivå følgende hypoteser: $H_0 : \theta = -10$ mot $H_1 : \theta < -10$. Hvilket av følgende resultater er da korrekt?

a) Vi får testobservator $T = -2.01$ og beholder nullhypotesen på 5% nivå.
b) Vi får testobservator $T = -2.92$ og beholder nullhypotesen på 5% nivå.
c) Vi får testobservator $T = -2.92$ og forkaster nullhypotesen på 5% nivå.
d) Vi får testobservator $T = -2.01$ og forkaster nullhypotesen på 5% nivå.
e) Vi får testobservator $T = 2.01$ og beholder nullhypotesen på 5% nivå.

Oppgave 21
Man laget også et konfidensintervall for en annen kontrast definert ved $\theta = \mu_3 - \mu_4$ og intervallet ble [-49.22, -2.78]. Hva er konfidensnivået på dette intervallet?

a) 80% b) 90% c) 95% d) 98% e) 99%

Oppgave 22
Determinasjonskoeffisienten R^2 er lik

a) 0.996 b) 0.532 c) 0.501 d) 0.843 e) 0.457

Oppgave 23
Hva er utvalgsstandardavviket for gård 3 (tallet er utelatt i R-utskriften)

a) 15.55 b) 15.16 c) 16.72 d) 19.23 e) 18.03

Sykehusutgifter (Oppgave 24-32)
I 2008 ble det samlet inn opplysninger fra *n*=10 norske sykehus om utgifter (målt i 1000 kr) og antall sykehussenger de hadde på henholdsvis kirurgisk og medisinsk avdeling. Nedenfor og på neste side finner du R Commander utskrifter som viser noe deskriptiv statistikk om variablene og resultatet av to ulike regresjonsanalyser herved kalt «Modell 1» og «Modell 2». Bruk utskriften i den grad du finner det nødvendig til å besvare spørsmålene.

```
                mean         sd   n
utgifter    102294.3   46674.91  10
medisinsk       58.4      30.15  10
kirurgisk       55.2      19.53  10
```

```
> summary(Modell 1)

Call:
lm(formula = utgifter ~ medisinsk, data = Sykehus)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19827.7    14581.0   1.360 0.210966
medisinsk     1412.1      224.2   6.298 0.000233

s: 20280 on 8 degrees of freedom
Multiple R-squared: 0.8322

> summary(Modell 2)

Call:
lm(formula = utgifter ~ kirurgisk, data = Sykehus)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -719.3    30725.8  -0.023  0.98190
kirurgisk     1866.2      527.7       *  0.00766

s: 30920 on 8 degrees of freedom
Multiple R-squared: 0.6099
```

Oppgave 24
Fra Modell 1: Hvor mye er den estimerte endringen i utgifter (målt i 1000kr) dersom antall senger på medisinsk avdeling økes med 10 senger?

a) 1412.1 b) -1412.1 c) 198277 d) 19827.7 e) 14121

Oppgave 25
Fra Modell 1: Finn et 95% konfidensintervall for effekten (β) av antall senger (i medisinsk avdeling) på sykehusutgifter.

a) [895.1, 1929.1]
b) [995.1, 1829.1]
c) [794.3, 1852.5]
d) [659.9, 2164.3]
e) [963.7, 1860.5]

Oppgave 26
Fra Modell 2: Hva vil du anslå utgiftene (i 1000 kr) til å være for et sykehus med 40 senger på kirurgisk avdeling?

a) 74211.8 b) 69333.7 c) 71234.4 d) 83122.1 e) 73928.7

Oppgave 27
Fra Modell 2: Hva blir et 95% prediksjonsintervall for utgiftene (i 1000 kr) til et vilkårlig sykehus med gjennomsnittlig antall senger på kirurgisk avdeling? Du kan her benytte at modellens anslag på utgiftene vil være $\hat{y} = 102295$.

a) [41977, 162613]
b) [27513, 177077]
c) [30043, 174547]
d) [102367, 102223]
e) [-6505, 211095]

Oppgave 28
Hva er korrekt tolkning av prediksjonsintervallet du skulle finne i oppgave 27?

a) Det er 95% sannsynlig at forventningen til utgiftene til et sykehus med gjennomsnittlig antall senger på kirurgisk avdeling vil ligge i dette intervallet
b) Det er 95% sannsynlig at intervallet dekker den sanne, ukjente forventningen til utgiftene til et vilkårlig sykehus med gjennomsnittlig antall senger på kirurgisk avdeling .
c) Det er 95% sannsynlig at de faktiske utgiftene til et vilkårlig sykehus med gjennomsnittlig antall senger på kirurgisk avdeling vil ligge i dette intervallet.
d) Det er 95% sannsynlig at det estimerte prediksjonsintervallet inneholder det faktiske antall senger på kirurgisk avdeling på det aktuelle sykehuset.
e) Det er 95% sannsynlig at det gjennomsnittlige utgiftsnivået for sykehus med gjennomsnittlig antall senger på kirurgisk avdeling ligger i dette intervallet.

Oppgave 29
I Modell 1 er følgende lineære modell antatt for sammenhengen mellom utgifter (*y*) og antall senger på medisinsk avdeling (*x*): $y_i = \alpha + \beta x_i + \epsilon_i$. Hvilken av følgende tilleggsantagelser har vi IKKE gjort i dette kurset?

a) Alle ϵ_i (for $i = 1, \ldots, n$) er uavhengige av hverandre
b) Støyleddet er normalfordelt
c) Populasjonsstandardavviket til ϵ_i er uavhengig av x
d) $E(y|x) = 0$
e) $y_i \sim N(\alpha + \beta x_i, \sigma)$

Oppgave 30
I Modell 1 ønsker vi å teste om det er en positiv lineær sammenheng mellom antall senger og utgifter. Hvilke hypoteser er da riktig for denne testen?

a) $H_0 : \beta = 0$ mot $H_1 : \beta > 0$
b) $H_0 : \beta = 0$ mot $H_1 : \beta < 0$
c) $H_0 : \alpha = 0$ mot $H_1 : \alpha > 0$
d) $H_0 : \alpha = 0$ og $\beta = 0$ mot $H_1 : \alpha > 0$ og $\beta > 0$
e) $H_0 : \beta = 0$ mot $H_1 : \beta \neq 0$

Oppgave 31
Basert på resultatene fra Modell 2, hva blir testobservatoren for testen om det er effekt av antall senger på utgifter til sykehus (tallet er fjernet fra utskriften).

a) 3.976 b) 3.012 c) 4.125 d) 3.245 e) 3.536

Oppgave 32
Basert på utskriften for de to modellene; Hva er den mest korrekte avslutning på følgende utsagn: «Av de to modellene gir Modell 1 best tilpasning til dataene i det undersøkte utgiftsintervallet fordi…»

a) …gjennomsnittlig antall senger på medisinsk avdeling er høyere enn på kirurgisk.
b) …det estimerte konstantleddet i Modell1 ikke er negativt.
c) …det er større variasjon i antall senger på medisinsk avdeling.
d) …den forklarer en større del av variasjonen i utgiftene.
e) …den forklarer en større del av variasjonen i antall senger.

Arbeidsundersøkelse (Oppgave 33-37)
I en arbeidsundersøkelse ville man se på sammenhengen mellom faktorene næring og yrke. Spesielt ville man se på fire næringer (industri, bygg/anlegg, varehandel og informasjon/kommunikasjon) og tre yrker (leder, kontor, salg/service). Resultatet av undersøkelsen der n=182 personer ble spurt, er gitt i tabellen nedenfor.

Næring \ Yrke	**Leder**	**Kontor**	**Salg/service**	Total
Industri	19	17	5	41
Bygg/anlegg	14	10	5	29
Varehandel	13	10	56	79
Info/kommunikasjon	18	11	4	33
Total	64	48	70	182

Oppgave 33
Under en nullhypotese om at næring og yrkesvalg er uavhengige variabler, hva er estimatet for forventet antall som jobber med salg/service i varehandelen?

a) 30.4 b) 25.3 c) 35.7 d) 12.7 e) 53.1

Oppgave 34
Hva blir under H_0 bidraget til den kji-kvadratfordelte testobservatoren Q fra kombinasjonen leder og varehandel. Her kan du benytte at forventet antall under H_0 (dvs E_{31}) er lik 27.8.

a) 5.2 b) 6.8 c) 7.9 d) 8.3 e) 9.5

Oppgave 35
Hvis du vil utføre en test på 1% nivå av hypotesen om uavhengighet mellom næring og yrke ved hjelp av en kji-kvadrattest, hva blir nedre grense for forkastningsområdet for testobservatoren Q?

a) 0.87 b) 9.21 c) 18.55 d) 16.81 e) 13.28

Oppgave 36

Det viste seg at det var klar avhengighet mellom næring og yrke ifølge en kji-kvadrattest, og fra dataene ser man at det særlig er næringen varehandel som skiller seg fra de andre. Derfor ble varehandel kuttet ut fra analysen for å se om det var avhengighet mellom de tre resterende næringer og yrke. En analyse i R gav følgende utskrift:

```
> .Table  # Counts
             Leder Kontor    Salg/service
Industri        19      17              5
Bygg/anlegg     14      10              5
Info/komm       18      11              4

> .Test$expected # Expected Counts
                 Leder    Kontor   Salg/service
Industri     20.30097 15.12621      5.572816
Bygg/anlegg 14.35922 10.69903      3.941748
Info/komm    16.33981 12.17476      4.485437

> round(.Test$residuals^2, 2) # Chi-square Components
             Leder Kontor   Salg/service
Industri      0.08   0.23          0.06
Bygg/anlegg   0.01   0.05          0.28
Info/komm     0.17   0.11          0.05
```

Anta et 5% testnivå i en test av hypotesene:

H_0 : Næring og yrke er uavhengige variabler

H_1 : Næring og yrke er avhengige variabler

Hvilket av følgende utsagn om testobservatoren Q *og* tilhørende konklusjon er da korrekt?

a) Q = 0.28 og næring og yrke anses som avhengige variabler
b) Q = 1.04 og næring og yrke anses som avhengige variabler
c) Q = 20.3 og næring og yrke anses som avhengige variabler
d) Q = 0.28 og næring og yrke anses som uavhengige variabler
e) Q = 1.04 og næring og yrke anses som uavhengige variabler

Oppgave 37

Hvorfor er gyldigheten av testen i foregående oppgave litt tvilsom?

a) Siden alle bidragene til testobservatoren er veldig små.
b) Fordi forventet antall under nullhypotesen er veldig små for yrket salg/service, spesielt for næringene bygg/anlegg og info/kommunikasjon.
c) Fordi et testnivå på 5% er alt for liberalt når antall frihetsgrader er lavt.
d) Siden antagelsen om kji-kvadratfordeling er tvilsom når antall frihetsgrader er mindre enn 5.
e) Siden servicenæringen info/kommunikasjon er svært forskjellig fra de to andre næringene som er mer industri/bygg-orienterte.

Fullføring av høyere utdanning (Oppgave 38-40)

Over de 13 årene fra år 2000 til 2012 ville man undersøke om prosentandelen (Y) av studenter ved høyere utdanning som fullfører utdanningen sin avhenger av utdanningsnivået til foreldrene. Man delte studentene inn i tre kategorier utfra foreldrenes utdanning: Ingen = Ingen av foreldrene har høyere utdanning, Kort = Minst én av

foreldrene har kortere (2-4 år) høyere utdanning, og Lang = Minst én av foreldrene har lang (>4år) høyere utdanning.
Observasjoner av prosentandel:

År	Ingen	Kort	Lang	
2000	32.8	28.3	24.7	
2001	32.2	29.2	23.5	
2002	31.7	28.9	23.5	
2003	30.3	29.9	23.8	
2004	30.5	31.6	24.7	
2005	29.5	30.2	23.5	
2006	29.7	31.0	22.9	
2007	29.3	31.7	22.9	
2008	28.6	32.4	22.6	
2009	27.4	32.1	22.6	
2010	27.9	31.5	22.1	
2011	26.8	32.2	22.3	
2012	26.4	32.2	21.8	Totalsnitt
Gj.snitt:	**29.47**	**30.86**	**23.15**	**27.83**

Følgende modell ble antatt for prosentandelen for fullført utdanning for foreldrekategori *i* (for *i*=1,2,3) i år *j* (for *j*=1,…,13):
$Y_{ij} = \mu_i + \epsilon_{ij}$ hvor $\epsilon_{ij} \sim N(0, \sigma)$ og alle ϵ_{ij} uavhengige av hverandre.
En variansanalyse ble kjørt i R Commander med følgende delvise utskrift:

```
> Anova(LinearModel)
             Df  Sum Sq Mean Sq  F value
ForeldreUtd   2       *       *        *
Residuals     *    83.1     2.3
```

Oppgave 38
Hvor mange frihetsgrader er knyttet til SS_E i denne modellen?
a) 39 b) 38 c) 36 d) 35 e) 33

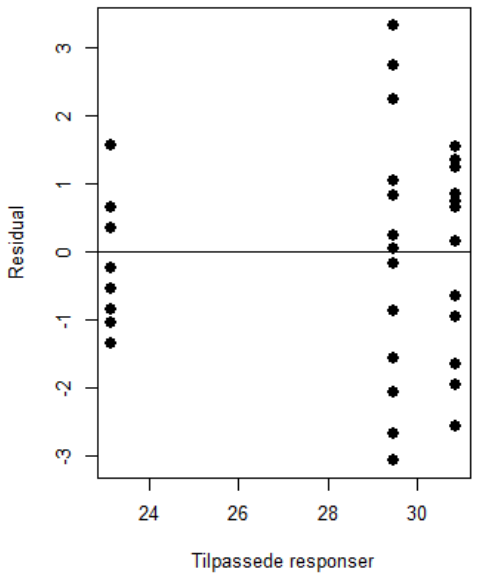
Oppgave 39
Hva er verdien av SS_G ?
a) 439 b) 166 c) 265 d) 578 e) 412

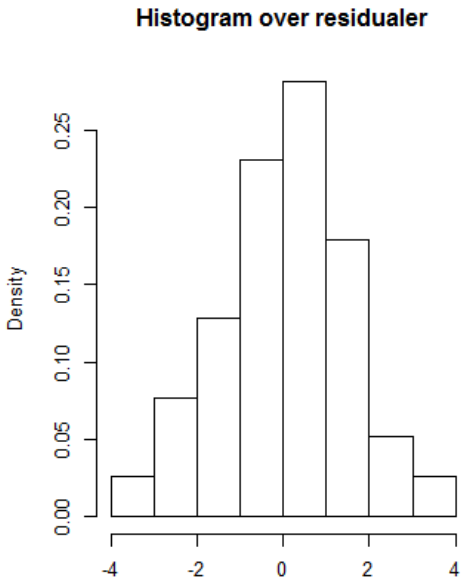
Oppgave 40
På neste side finner du tre residualplott som kan brukes til å vurdere om modellantagelsene er oppfylt. Figurene viser

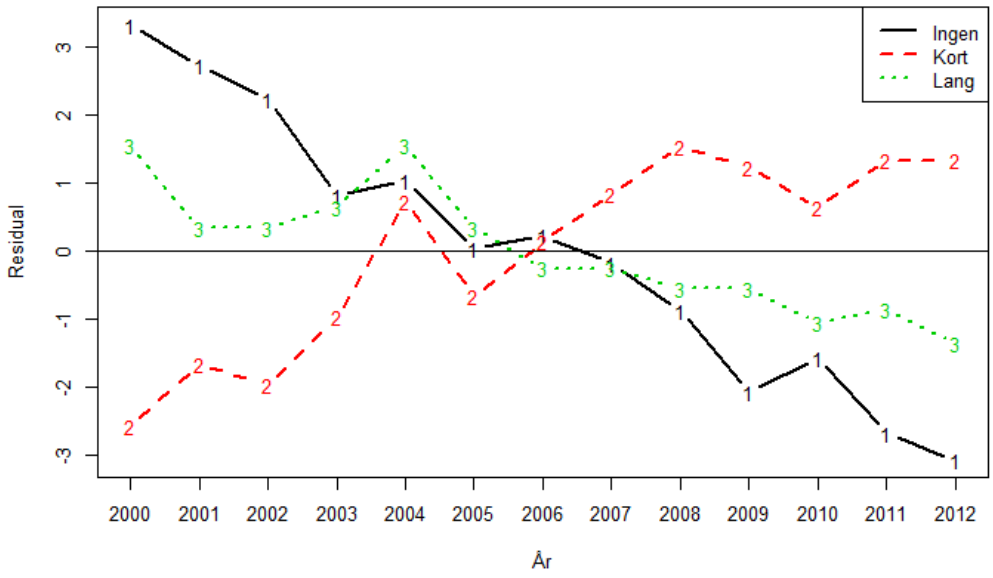
1) Residualer mot tilpassede responser ($\hat{Y}_{ij}$)
2) Histogram over residualer
3) Residualer plottet mot år for hver foreldrekategori

Hvilket par av modellantagelser er det størst grunn til å betvile er oppfylt på bakgrunn av disse plottene?

a) Antagelsene om normalfordelte og uavhengige støyledd.
b) Antagelsene om uavhengige støyledd og felles varians for alle tre grupper.
c) Antagelsene om at feilleddene har forventning 0 og at de er normalfordelte.
d) Antagelsene om lineær modell og normalfordeling for støyleddene.
e) Antagelsene om lineær modell og at feilleddene har forventning 0.







Emneansvarlig: Solve Sæbø

Sensor: Torfinn Torp

Vedlegg: F-tabell for $\alpha = 0.05$

F-Table Upper Tail Area of .05																				
		Numerator df																		
Denominator df		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	60	120	∞
	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	245.9	248.0	249.1	250.1	252.2	253.3	254.3
	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
	3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.8	8.8	8.8	8.8	8.7	8.7	8.7	8.6	8.6	8.6	8.5	8.5
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86	5.80	5.77	5.75	5.69	5.66	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62	4.56	4.53	4.50	4.43	4.40	4.37
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.84	3.81	3.74	3.70	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.41	3.38	3.30	3.27	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.12	3.08	3.01	2.97	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.90	2.86	2.79	2.75	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.74	2.70	2.62	2.58	2.54
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.61	2.57	2.49	2.45	2.40
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.51	2.47	2.38	2.34	2.30
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53	2.46	2.42	2.38	2.30	2.25	2.21
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46	2.39	2.35	2.31	2.22	2.18	2.13
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40	2.33	2.29	2.25	2.16	2.11	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35	2.28	2.24	2.19	2.11	2.06	2.01
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.31	2.23	2.19	2.15	2.06	2.01	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27	2.19	2.15	2.11	2.02	1.97	1.92
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.23	2.16	2.11	2.07	1.98	1.93	1.88
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.08	2.04	1.95	1.90	1.84
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.18	2.10	2.05	2.01	1.92	1.87	1.81
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15	2.07	2.03	1.98	1.89	1.84	1.78
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.13	2.05	2.01	1.96	1.86	1.81	1.76
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11	2.03	1.98	1.94	1.84	1.79	1.73
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.09	2.01	1.96	1.92	1.82	1.77	1.71
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.07	1.99	1.95	1.90	1.80	1.75	1.69
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.89	1.84	1.74	1.68	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92	1.84	1.79	1.74	1.64	1.58	1.51
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84	1.75	1.70	1.65	1.53	1.47	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.75	1.66	1.61	1.55	1.43	1.35	1.25
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.67	1.57	1.52	1.46	1.32	1.22	1.00

Kandidatnummer:_____

Svarskjema: Sett ett kryss i hver rad i den kolonnen som svarer til det alternativet du mener er riktig svar på spørsmålet. Det er kun tillatt å sette ett kryss i hver rad. (Dersom du vil endre svaret ditt, marker tydelig at du velger bort alternativet ved å skravere bort krysset.)

Oppgave	a	b	c	d	e
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					