EKSAMENSOPPGAVE

Institutt:	**IKBM**	
Eksamen i:	**STAT 100**	**Statistikk**
Tidspunkt for 10eksamen:	15. mai 2015	09.00-12.30. 3,5 timer
Kursansvarlig:	Trygve Almøy	

Tillatte hjelpemidler: C3. Alle typer kalkulatorer, alle andre hjelpemidler
Oppgaveteksten er på 11 sider, inkludert en tabell.

Oppgave 1 teller 40 % av denne eksamen, og alle 9 delspørsmål teller likt.
Flervalgsspørsmål teller 60 % av denne eksamen.

Merk; Side 11: Skjema fylles ut og arket leveres inn.

Oppgave 1

I en undersøkelse ble det trukket tilfeldig ut 18 forskjellige Fast Food restauranter (i USA) og fra disse ble det tatt prøver av tilfeldig valgte hamburgere, der en målte innhold av fett, protein og kilokalorier, alt målt pr 100 gram hamburger. Data kan du finne på slutten av oppgaven (Tabell 3).

Først ville en se på sammenhengen mellom fett og kalorier, og prøvde en lineær regresjonsmodell, der kalorier var respons og fett var forklaringsvariabel. Modellen ble analysert ved minste kvadraters metode. Se Tabell 1.

A) Sett opp regresjonsmodellen med de nødvendige antagelsene.
Tolk alle parameterne i modellen.
Estimer alle parameterne i modellen.

B) Finn R^2 og gi en forklaring på hva denne måler.

C) Lag et 95 % konfidensintervall for regresjonskoeffisienten (β).
Ernæringseksperter har funnet ut at ett gram fett gir ca. 9 kalorier. Er det en motsetning mellom dette utsagnet og konfidensintervallet (svaret må begrunnes)?

D) Hva er sammenhengen mellom et konfidensintervall og en tosidig test? Svar helst med hensyn til det intervallet du nettopp har konstruert.

E) En hamburger inneholder 16,6 gram fett, hva vil du anslå kalori-innholdet til?
Gi et 95 % prediksjonsintervall for dette anslaget (du får oppgitt at gjennomsnittlig fettmengde er 16,6 gram).
Hvorfor er det for akkurat denne fettmengden intervallet blir smalest?

F) Finn residualet for den første hamburgeren i Tabell 3 og gi en forklaring på hva denne verdien betyr.

Se på residualplottene (Figur 1 og Figur 2), forklar hvordan du kan bruke disse for å si noe om eventuelle avvik fra modellantagelsene?

En prøvde med innhold av protein som forklaringsvariabel og kjørte modellen:

$$\text{Kalorimengde} = \alpha + \beta * \text{proteinprosent} + e,$$

med de vanlige modellantagelsene, og fikk resultater som i Tabell 2.

G) Gi en eller flere årsaker grunner at det er umulig å bruke protein til å forklare variasjon i kalori-innhold? (Vi er ute etter en statistisk forklaring, men har du en ernæringsmessig forklaring så kom med den, det trekker i alle fall ikke karakteren ned).

Hvis $\beta = 0$, hva er $P(|\hat{\beta}| \geq 0{,}7815)$?

I tillegg spurte en tilfeldige kunder om å gi smak poeng fra 0 til 10 (med 10 som best) til de forskjellige hamburgerne. Deretter ble en regresjonsmodell med fett som forklaring og poeng som respons kjørt. Dette ga en estimert linje:

$$\text{Poeng} = 7{,}6 - 0{,}1 * \text{fett},$$

og et residualplott som vist i Figur 4.

H) Hvordan kan dette plottet være med på å støtte om en, men bare en av disse tre påstandene:

Folk liker magre hamburgere
Folk liker hamburgere med noe fett, men ikke for mye.
Folk liker hamburgere med mye fett.

Skisser et spredningsplot av poeng (respons) mot fett (forklaringsvariabel).
Forklar hvorfor en rett linje ikke er brukbar som modell for disse data.

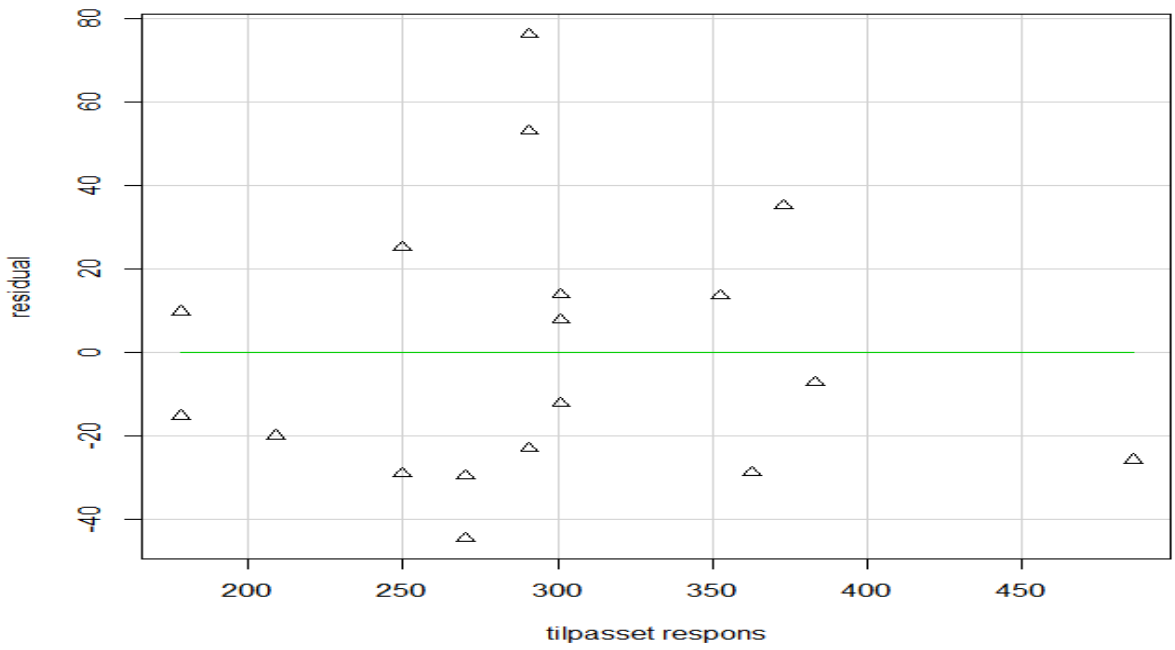
```
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  126.920     19.538   6.496 7.37e-06 ***
fett          10.250      1.082
s: 33.08 on 16 degrees of freedom

Analysis of Variance Table
          Df Sum Sq Mean Sq
fett       1  98172   98172
Residuals 16  17507    1094
```

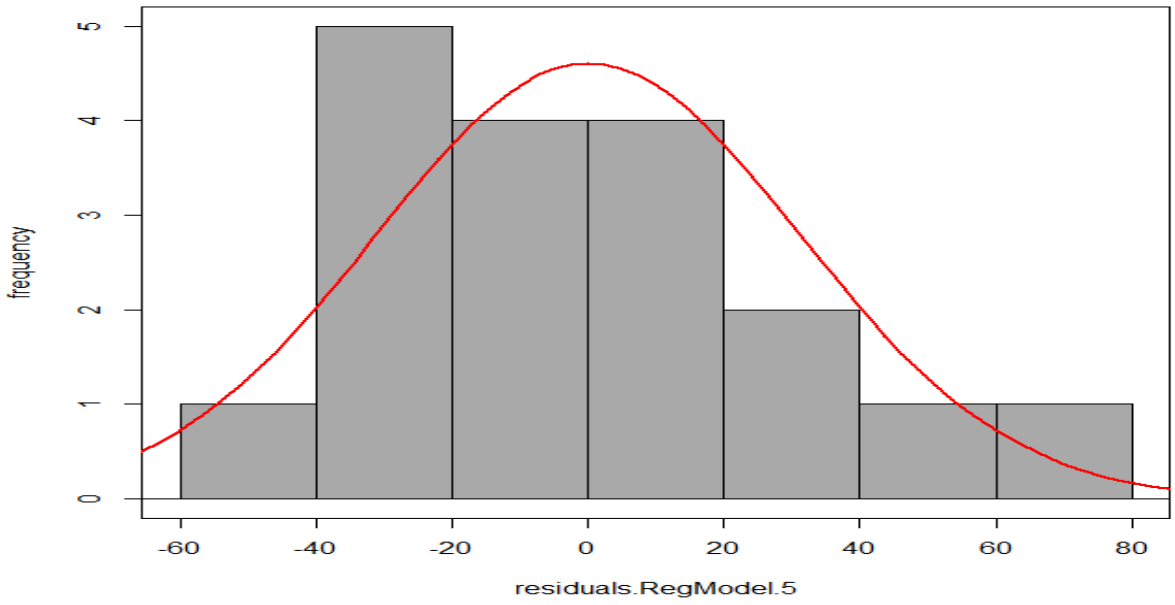
Tabell 1. Fett som forklaringsvariabel.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 305.6849    45.3009   6.748 4.68e-06 ***
protein      -0.7815     3.5003  -0.223    0.826
s: 84.9 on 16 degrees of freedom
Multiple R-squared: 0.003106,
```

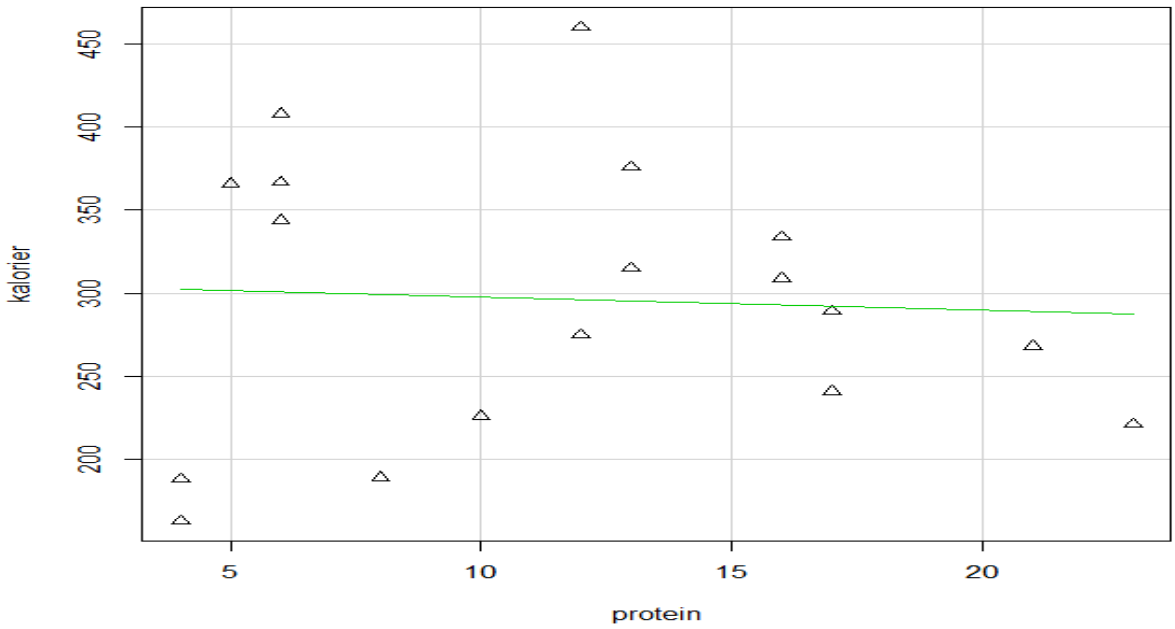
Tabell 2. Protein som forklaringsvariabel.



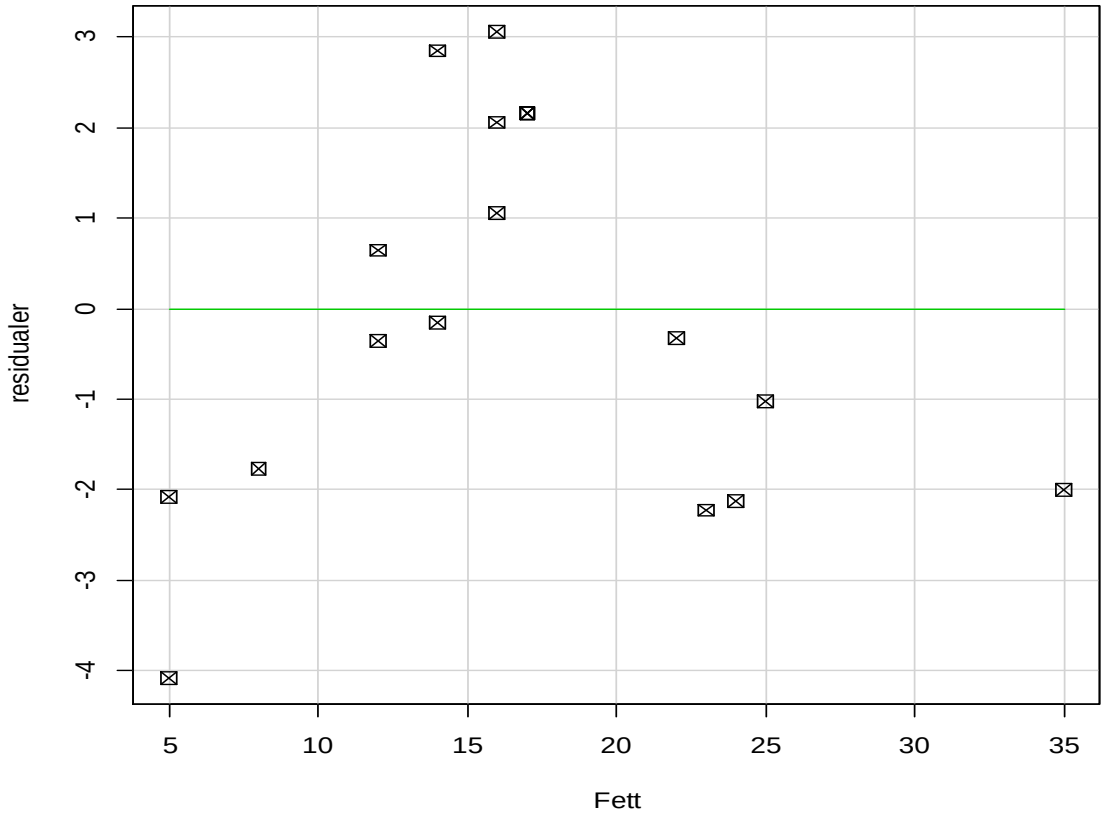
Figur 1. Residualer mot tilpassede verdier, fett er forklaringsvariabel



Figur 2, Histogram av residualer, fett er forklaringsvariabel.



Figur 3: Kalorimengde mot protein-innhold, med inntegnet tilpasset regresjonslinje.



Figur 4: Residualer mot tilpassede verdier. Respons er poeng, mens forklaringsvariabel er fett

Fett	Kalori	Protein	Poeng
17	289	17	8
25	376	13	4
17	315	13	8
16	268	21	7
12	221	23	7
17	309	16	8
14	226	10	6
16	344	6	9
5	163	4	3
16	367	6	8
5	188	4	5
12	275	12	6
24	408	6	3
35	460	12	2
8	189	8	5
22	366	5	5
23	334	16	3
14	241	17	9

Tabell 3: Data for oppgave 1

Flervalgsoppgaver

Dersom du får andre svar enn de som er oppgitt, kan det skyldes avrunding. Velg det alternativet som er nærmest.

For **oppgavene 1 - 5** anta at du har 5 observasjoner (X_1, X_2, X_3, X_4, X_5), alle er uavhengige og normalfordelte med forventning μ og standardavvik σ, der begge parameterne er ukjente.

La $\bar{X} = \frac{1}{5}\sum_{i=1}^{5} X_i$, og s = $\sqrt{\frac{1}{4}\sum_{i=1}^{5}(X_i - \bar{X})^2}$

Oppgave 1.

Dersom du estimerer μ med $\bar{X}$, hva blir standardfeilen til estimatoren?

A) 5 B) 4 C) $\frac{s}{5}$ D) $\frac{s}{\sqrt{5}}$ E) $\frac{\bar{X}}{\sqrt{5}}$ F) $\frac{s^2}{5}$

Oppgave 2.

Et 95 % konfidensintervall for μ er gitt ved:

A) $\bar{X} \pm 2{,}776\frac{s}{\sqrt{5}}$ B) $\bar{X} \pm 2{,}776\frac{s}{5}$ C) $\bar{X} \pm 2{,}776\frac{s}{\sqrt{4}}$

D) $\bar{X} \pm 2{,}776\frac{\sigma}{5}$ E) $\bar{X} \pm 2{,}571\frac{s}{\sqrt{5}}$ F) $\bar{X} \pm 2{,}571\frac{s}{\sqrt{4}}$

Oppgave 3.

Hvis du vil teste H_0: $\mu = 7$ mot H_1: $\mu > 7$ med signifikansnivå 5 %, bruker du regel:

A) Forkast H_0 hvis $\bar{X} > 7$ B) Forkast H_0 hvis $\mu > 7$

C) Forkast H_0 hvis $\bar{X} > 2{,}132$ D) Forkast H_0 hvis $\frac{\bar{X} - \mu}{s/\sqrt{5}} > 2{,}132$

E) Forkast H_0 hvis $\frac{\bar{X} - 7}{s/\sqrt{5}} > \mu$ F) Forkast H_0 hvis $\frac{\bar{X} - 7}{s/\sqrt{5}} > 2{,}132$

Oppgave 4.

Hvis P-verdien i testen i oppgave 3 blir på 0,03, vil du da:

A) Forkaste H_1. B) Forkaste H_0.

C) Ikke forkaste H_0. D) Gjøre en type 2 feil.

E) Påstå at sannsynligheten for H_0 er 0,03. F) Påstå at H_1 ikke kan være sann.

Oppgave 5.

Hvis du skal ha et forventningsrett estimat for variansen (σ^2) bruker du?

A) $\sqrt{\frac{1}{4}\sum_{i=1}^{5}(X_i - \bar{X})^2}$ B) $\frac{1}{4}\sum_{i=1}^{5}(X_i - \bar{X})^2$ C) $\left[\frac{1}{4}\sum_{i=1}^{5}(X_i - \bar{X})\right]^2$

D) $\frac{1}{5}\sum_{i=1}^{5}(X_i - \bar{X})^2$ E) $\frac{1}{4}\sum_{i=1}^{5}(X_i - \bar{X})$ F) $X_5 - X_1$

For oppgavene 6 – 8 anta at vi har målt proteininnhold i melk (målt i prosent) for 6 kyr av rasen Vanlig Norsk Ku (NRF) og 5 kyr fra en eldre rase, Sidet Trønder og Nordlandsfe (STN). Dyra er tilfeldig plukket ut, de sto i samme fjøs og fikk samme diett under forsøket. Vi antar at proteinprosent er normalfordelt og at standardavviket (for hele populasjonen) er det samme i begge gruppene.

Resultatet ble:

NRF:	3,2	3,3	3,1	3,5	3,5	3,4
STN:	3,5	3,6	3,7	3,2	3,9	

Følgende størrelser ble beregnet på bakgrunn av observasjonene ovenfor:

	Gjennomsnitt	standardavvik
NRF	3.33	0,16
STN	3,58	0,26

Oppgave 6.
Hva er estimert forventet forskjell i proteinprosent mellom SNT og NRF?
A) 3,33 B) 3.54 C) 0,23 D) 0,08 E) 0,22 F) 0,25

Oppgave 7
Hva er estimert felles standardavvik?
A) 0,18. B) 0,26 C) 0,21 D) 0,22 E) 0,04 F) 0,1

Oppgave 8
Vi ønsker å teste om det er forventet forskjell i proteinprosent mellom rasene. Dette ga en test observator (T) med tallverdi 1,93. Kan du da:
A) Påstå at rasene har forskjellig forventet proteinprosent med et signifikansnivå på 1%
B) Påstå at rasen har likt forventet proteinprosent med et signifikansnivå på 10 %.
C) Påstå at rasene har forskjellig forventet proteinprosent med et signifikansnivå på 5 %.
D) Påstå at rasene har forskjellig forventet proteinprosent uansett valg av signifikansnivå.
E) Påstå at rasene har likt forventet proteinprosent uansett valg av signifikansnivå.
F) Påstå at rasene har forskjellig forventet proteinprosent ved signifikansnivå på 10 %.

For oppgavene 9 - 15: I tillegg til forsøket beskrevet før oppgave 6 valgte en å måle proteinprosent på to utenlandske raser, (Holstein Frieser, HF og Finsk Ayrshire, FA). Kyr fra disse raser ble satt i samme fjøs som beskrevet tidligere og gitt samme diett. Disse ga følgende proteinprosenter:

HF:	3,2	3,2	3,1	3,0	3,4	3,5
FA	3,2	3,4	3,7	3,7	3,8	3,9

Det ble bestemt å analysere data fra alle 4 kurasene med en enveis variansanalysemodell (ONE WAY ANOVA)**.** Anta at alle observasjonene er uavhengige. La Y_{ij} være proteinprosent for ku nr. j tilhørende rase nr. i, der i = 1, 2, 3, 4, og j = 1, 2, ... , n_i.

R-commander ga følgende (redigerte) utskrift.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rase	3	0.6468			0.011
Residuals	19	0.8480			

```
     mean        sd       n
FA  3.63        0.23      6
HF  3.23        0.19      6
NRF 3.32        0.16      6
STN 3.58        0.26      5
```

Oppgave 9
Hvilken modell blir brukt?

A) $Y_{ij} = \mu_i + e_{ij}$, der $e_{ij} \sim N(0, \sigma)$.
B) $Y_{ij} = \mu_i + e_{ij}$, der $e_{ij} \sim N(0, \sigma_i)$.
C) $Y_{ij} = \mu + e_{ij}$, der $e_{ij} \sim N(0, \sigma)$.
D) $Y_{ij} = \mu_j + e_{ij}$, der $e_{ij} \sim N(0, \sigma)$.
E) $Y_{ij} = \mu_j + e_{ij}$, der $e_{ij} \sim N(0, \sigma_j)$.
F) $Y_{ij} = \mu_i + \sigma$.

Oppgave 10
Hva blir F verdien (F-value?)
A) 4,83 B) 0,66 C) 0,21 D) 0,22 E) 5,34 F) 2,20

Oppgave 11
Hva er det som blir testet i R utskriften?
A) H_0: Alle forventninger er like. H_1: Minst to forventninger er forskjellige.
B) H_0: Alle forventninger er like. H_1: Alle forventninger er forskjellig fra null.
C) H_0: Ingen forventninger er like. H_1: Alle forventninger er like.
D) H_0: Alle forventninger er null. H_1: Minst to forventninger er forskjellig fra null.
E) H_0: Samme varians i alle raser. H_1: Minst to rasevarianser er ulike.
F) H_0: Minst to forventninger er ulike. H_1: Alle forventninger er like.

Oppgave 12
Testen i Oppgave 11 har p-verdi på 0,011. Hva betyr det?
A) Sannsynlighet for at H_0 er rett dersom vi observerer så store raseforskjeller er 0,011.
B) Sannsynligheten for at H_0 er rett er 0,011.
C) Sannsynligheten for at H_0 er gal er 0,011.
D) H_0 kan ikke forkastes på 5 % signifikansnivå.
E) Sannsynligheten for å observere så store raseeffekter eller større er 0,011
F) Dersom H_0 er sann, er sannsynligheten for å observere så store eller større raseforskjeller lik 0,011.

Oppgave 13
Nummerer rasene slik: FA er 1, HF er 2, NRF er 3 og STN er 4. En kontrast (θ) som ser på forskjell mellom norske og utenlandske raser med hensyn på forventet proteininnhold er: $\theta = \frac{\mu_3+\mu_4}{2} - \frac{\mu_1+\mu_2}{2}$. Hva er et forventningsrett estimat for θ?
A) 0,02 B) 0,26 C) 0,40 D) 1 E) 13,76 F) 0,20

Oppgave 14
Et 95 % konfidensintervall for θ i oppgave 13 er gitt ved (-0.16; 0.21). Dette betyr at:
A) Vi kan påvise forskjell i forventet proteinprosent mellom de norske og de utenlandske rasene på 5 % signifikansnivå
B) Sannsynligheten for at det er forskjell i forventet proteinprosent mellom de norske og utenlandske rasene er 0,05
C) Intervallet gir ingen informasjon om proteininnholdet blant de 4 rasene.
D) Siden intervallet dekker null kan vi påstå at de norske har høyere proteinprosent.
E) Vi kan ikke påvise forskjell i forventet proteinprosent mellom de norske og de utenlandske rasene på 5 % signifikansnivå
F) Sannsynligheten for at det er forskjell mellom forventet proteininnhold er på minst 5 %.

Oppgave 15
Et estimat for variansen i proteinprosent innenfor hver rase (σ^2) er?
A) 4,83 B) 0,04 C) 0,21 D) 0,22 E) 0,848 F) 0,28

For oppgavene 16 og 17
Er merkeverksteder (M) dyrere enn ikke-merkeverksteder (IM)?
Vi har 5 biler som var lettere kollisjonsskadd. Hver bil ble tatt både til et M og et IM for å få et prisoverslag for å rette opp skadene. For hver bil ble det brukt nye verksteder.

Bil	1	2	3	4	5	1	2	3	4	5
Verkstedtype	M	M	M	M	M	IM	IM	IM	IM	IM
Prisoverslag (tusen kroner)	3	2,5	4	5	5	3	1,5	2	3	4

Oppgave 16
Hva blir testobservatoren (T) som brukes til å teste hypotesen om at M har høyere forventet pris (beskrevet som dyrere i svaralternativene) enn IM. Og hva er konklusjonen dersom du tester på signifikansnivå 5 %?
A) T = 1,43, vi kan ikke påstå at M er dyrere.
B) T = 1,43, vi kan påstå at M er dyrere.
C) T = 1,2, vi kan påstå påvis at M er dyrere.
D) T = 3,2, vi kan ikke påstå at M er dyrere.
E) T = 3,2, vi kan påstå at M er dyrere.
F) T = μ_d, vi kan påstå at M er dyrere.

Oppgave 17
En annet forslag til å teste om M er dyrere enn IM er H_0: p = 0,5 mot H_1: p > 0,5 der p er sannsynligheten for at M er dyrere enn IM for en vilkårlig kollisjonsskadd bil. Se bort fra det ene tilfelle der verkstedene var like dyre. Hva blir p-verdien for denne testen?
A) 0,05 B) 0,5 C) 1 D) 0 E) 0,9375 **F)** 0,0625

Følgende gjelder for **oppgavene 18-20**
Blod deles inn i 4 blodtyper (A, B, AB, og 0) som har å gjøre med variasjon i sukker i blodet, men blod kan også klassifiseres ved Rhesus (RH+ og RH-) som har med variasjon av

protein i blodet å gjøre. En undersøkelse av blod til 500 personer ga denne tabellen der vi har talt opp antallet innen hver kombinasjon:

	A	B	AB	0
RH+	176	28	22	198
RH-	30	12	4	30

Vi ønsker å teste om det er sammenheng mellom blodtype og Rhesus. En kjøring i R-commander ga (en redigert) utskrift som følger:

Expected counts:

	A	B	AB	0
RH+	174.7	33.9	22.0	193.3
RH-	31.3	**?**	4.0	34.7

Chi-square components:

	A	B	AB	0
RH+	0.01	1.03	**?**	0.11
RH-	0.05	5.76	**?**	0.63

Oppgave 18

Hva er forventet antall dersom det er uavhengighet mellom Blodtype og Rhesus for kombinasjonen **B** og **RH-**?

A) 5 B) 0,5 C) 0,011 D) 6,08 E) 0 F) 0,0625

Oppgave 19

Hva er Testobservatoren (kalt Q i læreboka) for denne testen?

A) 0,05 B) 7.58 C) 10,6 D) 5 E) 3 F) 0,0625

Oppgave 20

Testen fikk en p-verdi på 0,055. Kan du dermed forkaste H_0 om at det ikke er sammenheng mellom blodtype og Rhesus, på signifikansnivå 10 %, begrunnelsen må også være korrekt?

A) Ja, fordi Q er større enn 0,055.
B) Ja, fordi p-verdien er mindre enn 0,1.
C) Ja, fordi p-verdien betyr at H_1 er sann.
D) Nei, fordi p-verdien er mindre enn 0,1.
E) Nei, fordi p-verdien sier ikke noe om sannsynligheten for type 1 feil.
F) Nei, fordi p-verdien er større enn det som er vanlig i slike tester.

Riv ut arket og levere dette sammen med besvarelsen.
Bare ett kryss i hver rute.

Opp-gave	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						