

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

Eksamen i STK1000 — Innføring i anvendt statistikk.

Eksamensdag: Onsdag 7. oktober 2009.

Tid for eksamen: 15:00–17:00.

Oppgavesettet er på 8 sider.

Vedlegg: Ingen

Tillatte hjelpemidler: Lærebok, ordliste for STK1000, godkjent kalkulator.

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

Husk å fylle inn kandidatnummer under.

Kandidatnr: \_\_\_\_\_

Alle 20 oppgaver teller likt. For hver oppgave skal du merke av for bare ett svaralternativ. Du får ett poeng for hvert riktige svar, maksimum 20 poeng. Dersom du svarer feil eller lar være å krysse av på et spørsmål, får du null poeng. Du blir altså ikke “straffet” med minuspoeng for å svare feil. *Lykke til!*

**Oppgave 1.** Den diskrete tilfeldige variabelen  $X$  har sannsynlighetsfordeling

|              |      |     |       |
|--------------|------|-----|-------|
| $x_i$        | −100 | 0   | $y$   |
| $P(X = x_i)$ | 0.3  | 1/2 | $p_3$ |

hvor tallene  $y$  og  $p_3$  ikke er oppgitt. Da er

- a) det ikke mulig å si hva  $p_3$  er uten å vite verdien til  $y$   
 b)  $p_3 = 0.20$      c)  $p_3 = 0.30$      d)  $p_3 = 0.80$

**Oppgave 2.** Den diskrete tilfeldige variabelen  $X$  har sannsynlighetsfordeling

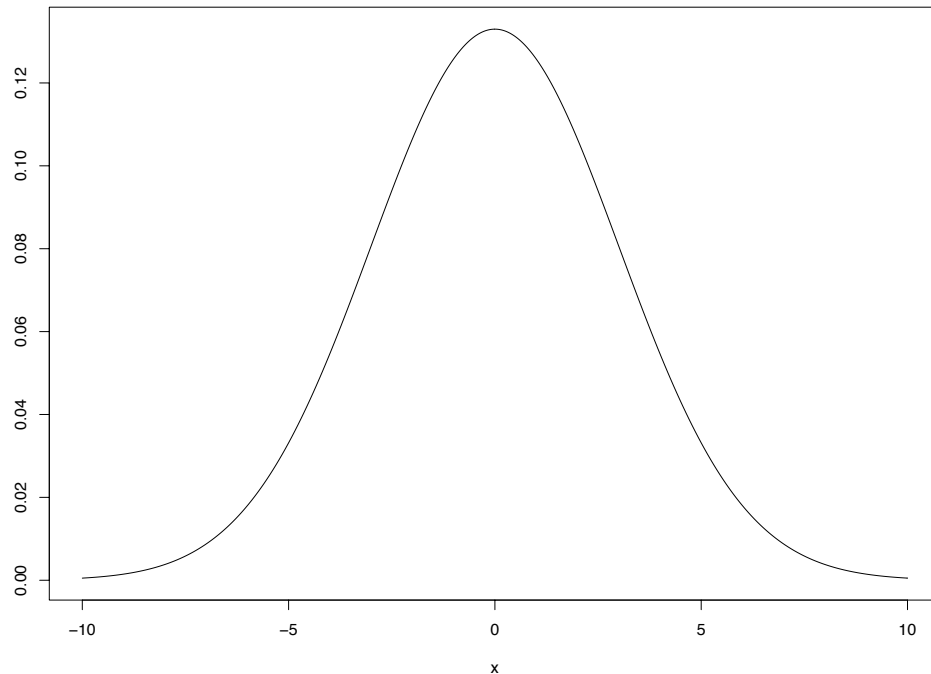
|              |     |     |     |
|--------------|-----|-----|-----|
| $x_i$        | −1  | 0   | $z$ |
| $P(X = x_i)$ | 1/6 | 2/3 | 1/6 |

hvor tallet  $z$  ikke er oppgitt. Forventningen til  $X$  er 1.5. Da må  $z$  være

- a) 1     b) 1.5     c) 2.5     d) 10

(Fortsettes på side 2.)

**Oppgave 3.** Under er tetthetskurven til en  $N(0, \sigma)$  fordelt variabel.



Uten noen annen informasjon enn tegningen over kan man se hva standardavviket  $\sigma$  skal være utifra STK1000-teori. Det er nemlig slik at  $\sigma$  er lik

- a) 1     b) 3     c) 5     d) 9

**Oppgave 4.** I følgende ordnete datasett med  $n = 8$  observasjoner har observasjonen merket med ? falt ut.

24, 56, 56.78, ?, 700, 1405, 1560.3, 2800

Hvis vi vet at medianen  $m = 650$ , vil den manglende observasjonen verdien

- a) være lik 700     b) være mellom 56.78 og 378.39  
 c) ikke kunne bestemmes     d) være lik 600

**Oppgave 5.** To tilfeldige variabler  $X$  og  $Y$  har standardavvik  $\sigma_X = 3$  og  $\sigma_Y = 1$ . Korrelasjonen mellom  $X$  og  $Y$  er  $-0.5$ . Da har  $X - Y$  standardavvik lik

- a) 3.606     b) 3.162     c) 13     d) 2

**Oppgave 6.** Temperaturmålinger hos en frisk person kan antas å være  $N(\mu, \sigma)$  med  $\mu = 37.0^\circ$  og  $\sigma = 0.3^\circ$ . Da er sannsynligheten for at en

(Fortsettes på side 3.)

måling av en frisk person viser mer enn  $37.5^\circ$

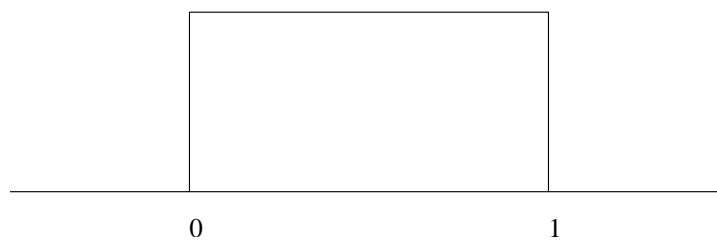
- a) 0.0764     b) 0.0344     c) 0.0475     d) 0.9525

**Oppgave 7.** Vi fortsetter med å jobbe med temperaturmålinger som i forrige oppgave. For en viss temperaturverdi er sannsynligheten omtrent 0.75 for at en temperaturmåling skal være mindre enn denne verdien.

Temperaturverdien er

- a)  $37.2^\circ$      b)  $37.4^\circ$      c)  $37.6^\circ$      d)  $40.0^\circ$

**Oppgave 8.** Under er tetthetskurven til en kontinuerlig tilfeldig variabel  $X$ .



Da er forventningen ( $\mu_X$ ) lik

- a) 0     b) 0.25     c) 0.5     d) 0.75

**Oppgave 9.** Anta man observerer tusen uavhengige normalfordelte variabler  $X_1, X_2, \dots, X_{1000}$  som alle har fordeling  $N(0, \sigma)$ . Da er gjennomsnittet  $\frac{1}{1000} \sum X_i$

- a) cirka  $1000\sigma$  siden man legger sammen tusen tall som har et typisk avvik på  $\sigma$   
 b) cirka 0     c) Binomisk fordelt med  $n = 1000$      d) cirka  $\sigma$

**Oppgave 10.** En kvinne tar en mammografiundersøkelse. La

$$S = \{\text{kvinnen har brystkreft}\}$$

og

$$M = \{\text{mammogrammet viser tegn på kreft}\}.$$

Anta at  $P(M|S) = 0.95$ ,  $P(M|S^c) = 0.035$  og  $P(S) = 0.007$ . Da er  $P(M)$  lik

- a) 0.00665     b) 0.0622     c) 0.16     d) 0.041405

**Oppgave 11.** En *polygraf*, det vil si en løgndetektormaskin, brukes ofte på ansatte i sensitive stillinger i USA. En slik maskin må kalibreres. Man stiller da forsøkspersoner, som man vet om lyver eller ikke, spørsmål og ser hvordan maskinen reagerer. Grensen for når maskinen sier at en person lyver er da fastsatt i forhold til dette.

Definer begivenhetene

$$M = \{\text{maskinen viser indikasjon på at man lyver}\}$$

(Fortsettes på side 4.)

og

$$L = \{\text{Personen lyver}\}.$$

En typisk ferdigkalibrert maskin har  $P(M|L) = 0.88$  og  $P(M|L^c) = 0.14$ . Det er dessuten rimelig å anta at flertallet man bruker løgndetektortesten på i en ansettelsesprosess vil snakke sant, og vi antar derfor  $P(L^c) = 0.99$ . Da er sannsynligheten for at man *ikke* lyver, gitt at løgndektoren påstår at man lyver, det vil si  $P(L^c|M)$ , lik

- a) 0.86     b) 0.94     c) 0.001     d) 0.0597

**Oppgave 12.** I et medisintisk studium målte man FGP ("Fasting plasma glucose") og HbA (som svarer ca til prosentandel røde blodlegemer som har glukose molekyler festet på seg) på 18 diabetespasienter. Under vises resultatet Minitab gir av en regresjonsanalyse til datasettet.

Regression Analysis: fpg versus hba

The regression equation is

$$\text{fpg} = 66,4 + 10,4 \text{ hba}$$

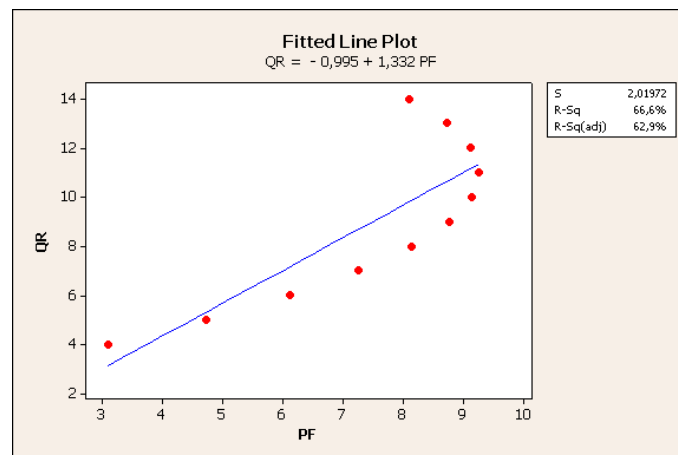
| Predictor | Coef   | SE Coef | T    | P     |
|-----------|--------|---------|------|-------|
| Constant  | 66,43  | 46,52   | 1,43 | 0,173 |
| hba       | 10,408 | 4,731   | 2,20 | 0,043 |

S = 63,8156    R-Sq = 23,2%    R-Sq(adj) = 18,4%

Da er stigningstallet til regresjonslinjen

- a) 66.4     b) 63.8156     c) 46.52     d) 10,4

**Oppgave 13.** Følgende datasett er konstruert for anledningen. Dette konstruerte datasettet er utgjort av to variabler som er kodet under navnene *PF* og *QR* og har 11 observasjoner. Minitab gir ut følgende regresjonsanalyse og plott av datasettet sammen med regresjonslinjen.



(Fortsettes på side 5.)

## Regression Analysis: QR versus PF

The regression equation is

$$QR = -0,99 + 1,33 PF$$

| Predictor | Coef   | SE Coef | T     | P     |
|-----------|--------|---------|-------|-------|
| Constant  | -0,995 | 2,435   | -0,41 | 0,692 |
| PF        | 1,3325 | 0,3144  | 4,24  | 0,002 |

S = 2,01972    R-Sq = 66,6%    R-Sq(adj) = 62,9%

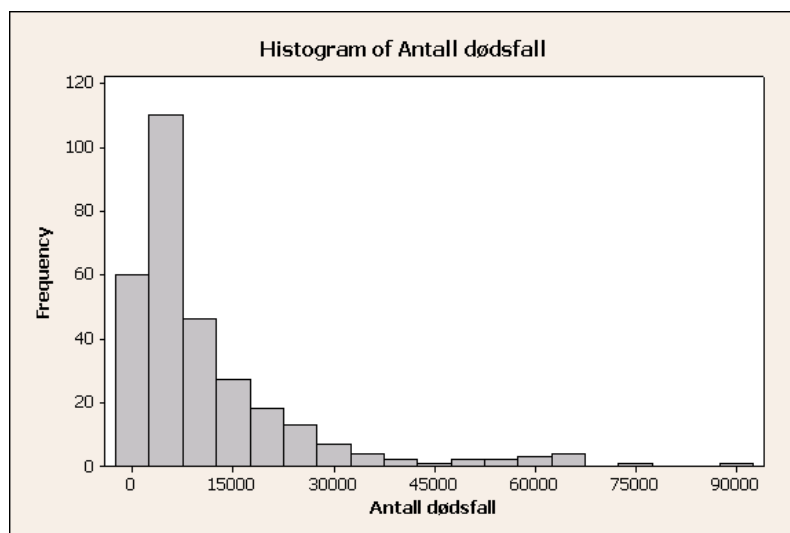
Under er en liste med fire utsagn, hvor av bare ett er feil. Kryss av på hvilket utsagn som ikke er sant.

- a) Regresjonslinjen skjærer  $y$ -aksen i  $-0.99$  og har et positivt stigningstall
- b) Det er en klar lineær sammenheng i datasettet
- c) Korrelasjonen mellom  $QR$  og  $PF$  er  $0.816$
- d) Jo høyere  $PF$  er, jo høyere er typisk også  $QR$

**Oppgave 14.** I en fordeling som er skjev mot høyre er typisk

- a) medianen mindre enn gjennomsnittet
- b) mange observasjoner feilregistreringer
- c) medianen nesten lik gjennomsnittet hvis standardavviket er stort nok
- d) medianen er større enn gjennomsnittet

**Oppgave 15.** Under er et histogram over antall dødsfall i 301 forskjellige fylker i USA forårsaket av brystkreft hos kvinner i 1960.



Histogrammet er

- a) tilnærmet normalfordelt     b) tydelig høyreskjev
- c) tydelig venstreskjev     d) tydelig flertoppet

(Fortsettes på side 6.)

**Oppgave 16.** Her er oppsummerende statistikk av datasettet om antall dødsfall av brystkreft fra forrige oppgave .

Descriptive Statistics: Antall dødsfall

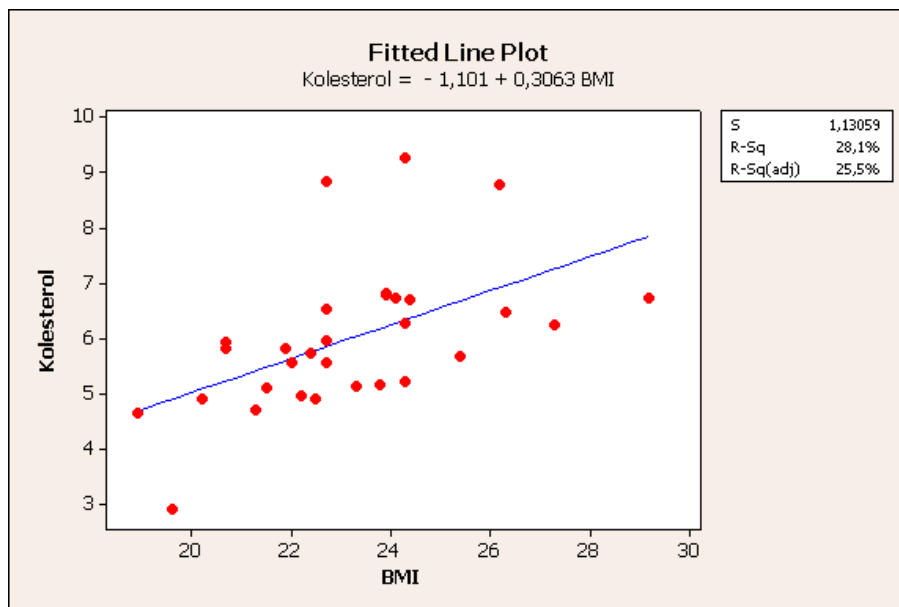
| Variable        | N   | N* | Mean  | SE Mean | StDev | Minimum | Q1   |
|-----------------|-----|----|-------|---------|-------|---------|------|
| Antall dødsfall | 301 | 0  | 11288 | 794     | 13780 | 445     | 2932 |

| Variable        | Median | Q3    | Maximum |
|-----------------|--------|-------|---------|
| Antall dødsfall | 6445   | 14039 | 88456   |

Da er IQR

a) 11107       b) 2932       c) 14039       d) 13780

**Oppgave 17.** En medisinsk undersøkelse målte kolesterol i blodet og BMI til 30 tilfeldig valgte kvinner. BMI står for body mass index regnes ut ved å ta vekt delt på høyde opphøyd i annen, hvor man måler vekt i kilogram og høyde i meter. Under følger noen Minitabutskrifter av en regresjonsanalyse som ble gjort for å undersøke sammenhengen mellom kolesterol og BMI.



Regression Analysis: Kolesterol versus BMI

The regression equation is

Kolesterol = - 1,10 + 0,306 BMI

| Predictor | Coef    | SE Coef | T     | P     |
|-----------|---------|---------|-------|-------|
| Constant  | -1,101  | 2,156   | -0,51 | 0,614 |
| BMI       | 0,30629 | 0,09256 | 3,31  | 0,003 |

S = 1,13059    R-Sq = 28,1%    R-Sq(adj) = 25,5%

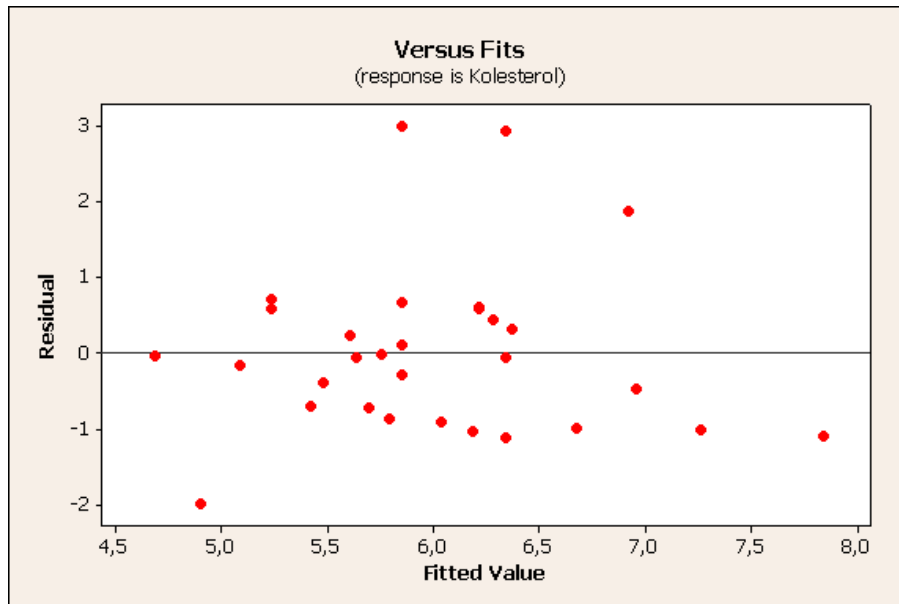
Unusual Observations

(Fortsettes på side 7.)

| Obs | BMI  | Kolesterol | Fit   | SE Fit | Residual | St Resid |
|-----|------|------------|-------|--------|----------|----------|
| 10  | 29,2 | 6,740      | 7,843 | 0,594  | -1,103   | -1,15 X  |
| 17  | 22,7 | 8,830      | 5,852 | 0,211  | 2,978    | 2,68R    |
| 25  | 24,3 | 9,270      | 6,342 | 0,231  | 2,928    | 2,65R    |

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.



Under følger fire utsagn, hvor *bare ett av dem er feil*. Her skal du altså krysse av på det ene svaralternativet som *ikke er riktig*.

- a) Kolesterol og BMI har en positiv sammenheng
- b) Residualene har ikke et veldig tydelig mønster
- c)  $r^2 = 28.1\%$  viser at mye variasjon ikke blir forklart
- d) Minitab skriver ut at man har innflytelsesrike observasjoner ("observation whose X value gives it large leverage"). Man kan derfor ikke si noe om sammenhengen mellom BMI og kolesterol.

**Oppgave 18.** Her er oppsummerende statistikk for BMI til datasettet i forrige oppgave.

Descriptive Statistics: BMI; Kolesterol

| Variable | N  | N* | Mean   | SE Mean | StDev | Minimum | Q1     |
|----------|----|----|--------|---------|-------|---------|--------|
| BMI      | 30 | 0  | 23,180 | 0,414   | 2,268 | 18,900  | 21,800 |

| Variable | Median | Q3     | Maximum |
|----------|--------|--------|---------|
| BMI      | 22,700 | 24,300 | 29,200  |

Da er variansen til BMI lik

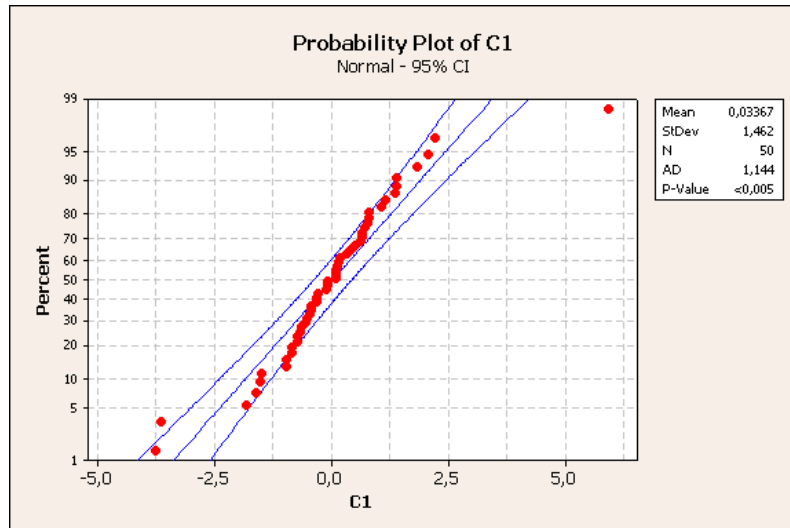
- a) ca 5.14     b) ca 0.173     c) ca 2.27     d) ca 0.28

(Fortsettes på side 8.)

**Oppgave 19.** Vi fortsetter med datasettet fra de to siste oppgavene. Fra informasjonen i utskriftene til de to siste oppgavene er *standardavviket* (det vil her si  $s_y$ ) til kolesterolobservasjonene

- a) ca 3.929     b) ca 1.31  
 c) ubestemmelig fra opplysningene i utskriftene     d) ca 2.47

**Oppgave 20.** Under er et kvantilplott som sammenlikner et datasett med normalfordelingen.



Man kan da si at

- a) observasjonene må være normalfordelte  
 b) observasjonene er tydelig flertoppet  
 c) det er grunn til å tvile på at observasjonene er normalfordelte  
 d) det er for få observasjoner ( $N = 50$ ) til å bruke et kvantilplott

*Det var det!*