EKSAMENSOPPGAVER

Institutt:	**IKBM**	
Eksamen i:	**STAT100**	**STATISTIKK**
Tid:	Torsdag 13.des 2012	09.00-12.30 (3.5 timer)
Emneansvarlig:	Solve Sæbø (90065281)	

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

Oppgaveteksten er på: 11

antall sider inkl. vedlegg

Alle deloppgaver teller likt. For hver oppgave er det 5 svaralternativer. Kun ett svaralternativ 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 11 og denne siden er den eneste som skal leveres når eksamen er slutt. Husk å skrive kandidatnummer på svarskjemaet!

Hastighet på biler (Oppgave 1-4)

På en vegstrekning har det i lang tid vært en fartsgrense på 50km/t. Fartsovervåkning av strekningen over svært lang tid indikerer at bilers hastighet (X) er normalfordelt med forventning μ=52km/t og et standardavvik på σ=10km/t. Hastigheten på strekningen ble så satt opp til 60 km/t. Noen dager etter fartshevningen ble hastigheten (Y) til n=5 tilfeldig valgte biler målt til:

63, 58, 45, 70, 66

Oppgave 1

Hva er et forventningsrett estimat på forventet fart etter fartshevningen?

a) 69.2 b) 61.3 c) 60.4 d) 62.7 e) 56.2

Oppgave 2
Du får i oppgave å teste om hevingen fra 50 til 60 km/t har ført til en forventet økning i fart på strekningen og synes 5 biler er for lite i utvalget. Du måler derfor farten til 16 nye biler og finner at gjennomsnittsfarten i det nye utvalget er $\bar{x}$=61.8 og at standardavviket er *s*=11.6. Hva blir verdien på den t-fordelte testobservatoren for denne testen (basert på kun det nye utvalget)?

a) 3.38 b) 21.3 c) 0.84 d) 3.92 e) 1.96

Oppgave 3
Vegvesenet har en mistanke om at folk venner seg til nye fartsgrenser og etter hvert endrer fartsvaner. Ett år etter at fartshevningen ble innført observerte man på nytt farten til 10 tilfeldig valgte biler. Gjennomsnittsfarten denne gangen ble 67.2 km/t og standardavviket ble 8.2 km/t. De ville sammenlikne de to utvalgene og antok at det ukjente populasjonsstandardavviket σ var det samme for begge år. Hva blir s_{pooled}, dvs fellesestimatet for σ basert på begge utvalgene på henholdsvis 16 og 10 biler?

a) 11.6 b) 109.3 c) 10.0 d) 9.9 e) 10.5

Oppgave 4
Vegvesenet definerte den tilfeldige variabelen *D* = *Y* – *Z*, der *Y* er fart rett etter fartshevningen til 60km/t og *Z* er farten ett år senere. Man antok at $D \sim N(\mu_D, \sigma_D)$. De formulerte følgende hypoteser: $H_0 : \mu_D = 0$ mot $H_1 : \mu_D < 0$ og testet hypotesen med 10% testnivå. Det førte til at nullhypotesen ble forkastet. Hva blir den praktiske konklusjonen av dette?

a) Den observerte farten ett år etter var i gjennomsnitt lavere enn før.
b) Den forventede farten hadde økt i løpet av året etter fartshevningen.
c) Det var ingen forskjell i forventet fart ett år etter fartshevningen.
d) Den forventede farten hadde gått ned i løpet av året etter fartshevningen.
e) Umulig å si noe siden utvalgene er så små.

Inntektsnivå hos kunder av ulike mobilselskap (Oppgave 5-10)
I en undersøkelse har man samlet inn data om inntektsnivået til 21 kunder av tre mobilfirmaer. I denne oppgaven skal vi studere disse dataene, og R utskriften i Tabell 1 nedenfor gir en del informasjon om disse. Merk at noen tall er utelatt fra utskriften.

Tabell 1

```
          mean        sd       n
Firma1    342.5     130.6      8
Firma2    167.1     113.5      7
Firma3    308.3      80.8      6

            Df Sum Sq Mean Sq F value
Firma        * 124538       *   4.886
Residuals   18 229376       *
Total        *      *
```

Oppgave 5
Hvor mange frihetsgrader har SSG (Sum of Squares Group), dvs kvadratsummen for faktoren Firma?

a) 21 b) 18 c) 20 d) 2 e) 3

Oppgave 6
Hvor mange frihetsgrader har SST (Sum of Squares Total)?

a) 21 b) 18 c) 20 d) 2 e) 3

Oppgave 7
Finn MSE (Mean Sum of Squares Error), dvs middelkvadratsummen knyttet til feilleddet.

a) 12743 b) 62269 c) 229376 d) 4128768 e) 124538

Oppgave 8
Dersom det hadde vært 7 observasjoner fra hvert firma, men ellers uendrede tall i Tabell 1, hva ville da MSE vært lik? (Husk at MSE også er estimatoren for felles varians innenfor alle grupper)

a) Uendret fra Tabell 1 b) 12657.07 c) 112.50 d) 12155.75 e) 110.25

Oppgave 9
Hva kalles vanligvis *sannsynligheten for å få en F-verdi større enn observert F=4.886* (under nullhypotesen om at det ikke er forskjell mellom firmaene med hensyn til forventet inntektsnivå til deres kunder)?

a) Testnivået α
b) Sannsynligheten for type II feil.
c) Sannsynligheten for at nullhypotesen er riktig.
d) Sannsynligheten for at alternativ hypotese er riktig.
e) p-verdien for testen.

Oppgave 10
Man ønsket å sammenlikne Firma 1 og Firma 3 ved å definere kontrasten
$\theta = \mu_1 - \mu_3$
Utfra resultatene i analysen i R fant man $\hat{\theta} = 34.17$ og $SE(\hat{\theta}) = 60.96$. Da vil et 90% konfidensintervall for θ være lik:

a) [-93.8, 162.2] b) [-71.5, 139.9] c) [11.2, 57.2] d) [-105.5, 105.5] e) [-26.8, 95.1]

Om effekt av pH på spiringsevne hos furufrø (Oppgave 11-20)
I en undersøkelse ble furufrø sådd i potter med fire ulike pH-nivåer: pH=3.8, pH=4.0, pH=4.2 og pH=4.4. (Vi betrakter i denne oppgaven pH nivå som en *kategorisk* variabel med disse fire kategoriene.) Ved hver pH ble det sådd frø i 5 potter og spiringsprosenten (Y) ble observert for hver potte. Forsøket gav følgende resultat:

pH 3.8	pH 4.0	pH 4.2	pH 4.4
73	83	87	93
83	80	97	97
67	90	90	83
62	77	90	91
80	85	86	91

R Commander gir følgende som utskrift

```
          mean   sd      n
pH 3.8    73   8.75      5
pH 4.0    83   4.95      5
pH 4.2    90   4.30      5
pH 4.4    91   5.10      5

             Df Sum Sq Mean Sq F value
pH            3   1034   344.6       *
Residuals    16    582    36.4
```

Oppgave 11
Hvilken modell er kjørt i henhold til R Commander utskriften? (For alle modeller antas alle ledd på høyre side av likhetstegnet å være uavhengige av hverandre.)

a) $Y_{ij} = \mu_i + \epsilon_{ij}$, der $\epsilon_{ij} \sim N(0, \sigma)$
b) $Y_{ij} = \mu + \epsilon_{ij}$, der $\epsilon_{ij} \sim N(0, \sigma)$
c) $Y_{ij} = \alpha + \beta + \epsilon_{ij}$, der $\epsilon_{ij} \sim N(0, \sigma)$
d) $Y_{ij} = \mu + \epsilon_{ij}$, der $\epsilon_{ij} \sim Bin(p)$
e) $Y_{ij} = \mu_i$, der $\mu_i \sim N(0, \sigma)$

Oppgave 12
Hva er verdien på F, dvs den observerte testobservatoren for testen om det er effekt av pH-nivå på forventet spiringsprosent?

a) 0.105 b) 1.78 c) 3.24 d) 5.29 e) 9.47

Oppgave 13
Hvor stor er R^2 ?

a) 0.64 b) 0.27 c) 0.36 d) 1 e) 0.80

Oppgave 14
Hvordan tolker vi determinasjonskoeffisienten R^2 i denne analysen?

a) Dersom den er negativ betyr det at modellen passer dårlig til data
b) Den er et mål på hvor mye av variasjonen i spiringsprosent som kan forklares ved modellen.
c) Den er et mål på variasjonen i spiringsprosent innen et gitt nivå av pH.
d) Den er et estimat på forventet spiringsprosent av furufrø basert på alle pH-grupper.
e) Den er et mål på hvor mye av variasjonen i pH som kan forklares fra spiringsprosenten.

Oppgave 15
Før forsøket hadde man en mistanke om at spireprosenten er forholdsvis god så lenge pH er minst 4.0. Følgende kontrast ble definert for å teste om det er noe generell forskjell i forventet spireprosent for pH lavere enn 4.0 og høyere enn eller lik 4.0:
$\theta = (\mu_2 + \mu_3 + \mu_4)/3 - \mu_1$
Der μ_1 er forventet spireprosent for pH=3.8, μ_2 for pH=4.0, osv. Hva er et forventningsrett estimat for denne kontrasten?

a) -9 b) -15 c) 15 d) 35 e) 10

Oppgave 16
Hva er standardfeilen til den estimerte kontrasten?

a) 1.03 b) 36.4 c) -3.97 d) 3.97 e) 3.12

Oppgave 17
Hvilke hypoteser er riktige å bruke for testen som er beskrevet i oppgave 15?

a) $H_0 : \theta = 0 \text{ mot } H_1 : \theta \neq 0$
b) $H_0 : \theta = 0 \text{ mot } H_1 : \theta > 3.8$
c) $H_0 : \theta = 0 \text{ mot } H_1 : \theta > 0$
d) $H_0 : \theta = 0 \text{ mot } H_1 : \theta < 0$
e) $H_0 : \theta = 3.8 \text{ mot } H_1 : \theta > 4.0$

Oppgave 18

Formelen for en kontrast i variansanalyse er generelt gitt ved: $\theta = \sum_{i=1}^{k} c_i \mu_i$

der *k* er antall grupper og c'ene er kjente konstanter. Anta at en kontrast er gitt på formen: $\theta = \mu_3 - (\mu_2 + \mu_4)/2$
Hvilke verdier har c_1, c_2, c_3 og c_4 dersom vi skal skrive den på den generelle formen (*k*=4)?

a) $c_1 = 1, c_2 = -1, c_3 = 1, c_4 = 0$
b) $c_1 = 0, c_2 = -1, c_3 = 1, c_4 = -1$
c) $c_1 = 0, c_2 = -1/2, c_3 = 1, c_4 = -1/2$
d) $c_1 = 1, c_2 = -1/2, c_3 = -1/2, c_4 = 0$
e) Den kan ikke skrives på den generelle formen

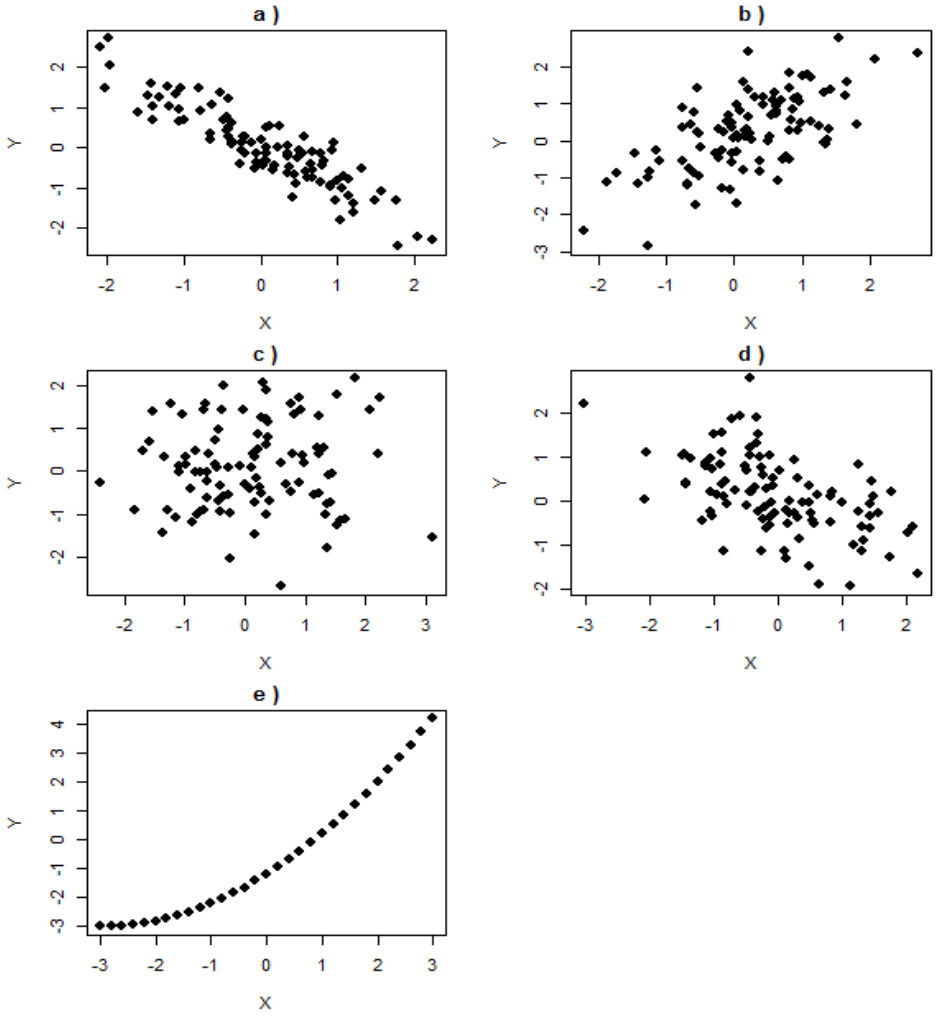
Oppgave 19
For generelle ANOVA modeller som vi har studert i dette kurset, hvordan bør vi tolke populasjonsstandardavviket σ?

a) Som et mål på variasjon mellom målinger av responsen gjort ved forskjellige nivåer av gruppe-variabelen.
b) Som et mål på totalvariasjon i spiringsprosent uavhengig av pH-nivå.
c) Som sannsynligheten for at målingen av responsen er korrekt.
d) Som et mål på variasjon mellom målinger av responsen gjort innen samme nivå av gruppe-variabelen.
e) Som forskjellen i forventet spiringsprosent mellom to ulike pH-verdier.

Oppgave 20
Anta ANOVA modellen $y_{ij} = \mu_i + \epsilon_{ij}$, der $\epsilon_{ij} \sim N(0, \sigma)$ og $i = 1, \ldots, k$ og $j = 1, \ldots, n_i$.
Hva er kaller vi da følgende størrelse: $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$?

a) SS_E b) R^2 c) SS_G d) SS_T e) MS_E

Om korrelasjon mellom to variabler (Oppgave 21-22)



Figur 1: Plott av ulike typer sammenhenger mellom X og Y

Oppgave 21

I Figur 1 er det plottet ulike sammenhenger mellom to variabler, X og Y. I hvilken delfigur er korrelasjonen mellom X og Y nærmest 0?

a) Figur a b) Figur b c) Figur c d) Figur d e) Figur e

Oppgave 22

Også med referanse til Figur 1. I hvilken delfigur er det størst negativ korrelasjon mellom X og Y?

a) Figur a b) Figur b c) Figur c d) Figur d e) Figur e

Sammenhengen mellom alder og maksimal puls (Oppgave 23-33)
I en undersøkelse ville man se på sammenhengen mellom alder (X) og maxpuls (Y) hos mennesker. Det ble trukket et tilfeldig utvalg på 10 personer som gjennomgikk en belastningstest på tredemølle og man målte makspuls på hver enkelt. Resultatene fra undersøkelsen og en utskrift fra en regresjonsanalyse er gitt nedenfor. Vi antar følgende modell: $y_i = \alpha + \beta x_i + \epsilon_i$, der $\epsilon_i \sim N(0, \sigma)$ og alle støyledd antas uavhengige for alle *i*=1,…10.

Maxpuls	Alder
186	30
183	38
171	41
177	38
191	29
177	39
175	46
176	41
171	42
196	24

```
Coefficients:
            Estimate Std. Error t value
(Intercept) 222.2707     6.0254       *
Alder        -1.1405     0.1612       *

s: 3.326 on * degrees of freedom
Multiple R-squared: 0.8622,
Adjusted R-squared: 0.8449
F-statistic: 50.04 on 1 and 8 DF,  p-value: 0.0001046
```

Oppgave 23
Hvor stor er den estimerte forventede endringen i Maxpuls dersom alderen øker med 10 år?
a) -1.1405 b) 222.27 c) 11.405 d) -11.405 e) 33.26

Oppgave 24
Hvor stor er den t-fordelte testobservatoren for å teste om det er lineær sammenheng mellom Alder og Maxpuls?
a) 36.889 b) -1.1405 c) -7.075 d) 0.8622 e) 0.1612

Oppgave 25
Hvor mange frihetsgrader har den t-fordelte testobservatoren for å teste hypotesen $H_0 : \beta = 0$ mot $H_0 : \beta \neq 0$?
a) 1 b) 8 c) 9 d) 7 e) 2

Oppgave 26
Noen hevder at maxpulsen har en forventet nedgang på ett slag i minuttet for hvert år man blir eldre. Da er det naturlig å test følgende hypoteser: $H_0 : \beta = -1$ mot $H_1 : \beta \neq -1$. Hva blir verdien for testobservatoren for denne testen?
a) -0.87 b) -15.48 c) -6.20 d) -7.074 e) 1.10

Oppgave 27
Dersom du heller skulle teste om forventet nedgang i maxpuls er på mer enn ett slag i minuttet for hvert år man blir eldre. Hva ville verdien av testobservatoren bli da?

a) Samme som i forrige oppgave. b) -14.15 c) -1.74 d) -0.43 e) -3.10

Oppgave 28
Et 95% konfidensintervall for β er lik:

a) [-1.68, -0.60] b) [-1.44, -0.84] c) [208.4, 236.2] d) [-1.46, -0.82] e) [-1.51, -0.77]

Oppgave 29
I undersøkelsen fant man et konfidensintervall for konstantleddet α lik [211.1, 233.5]. Hva er konfidensnivået for dette intervallet?

a) 85% b) 80% c) 95% d) 90% e) 99%

Oppgave 30
En mann ønsker å bruke den tilpassede modellen for å anslå sin egen maxpuls. Han er 30 år gammel. Hva blir den predikerte maxpulsen hans ifølge modellen?

a) 186.0 b) 177.3 c) 188.1 d) 256.5 e) 190.2

Oppgave 31
Hvor stort er residualet til den siste personen i undersøkelsen, han som er 24 år og har en maxpuls på 196?

a) 1.10 b) 0 c) -1.10 d) 1.21 e) -1.21

Oppgave 32
Gjennomsnittsalderen blant de som var med i undersøkelsen var 36.8 år. Hva blir et 95% prediksjons-intervall for maxpulsen til en vilkårlig person som er 36.8 år gammel?

a) [168.6, 192.0] b) [153.5, 207.1] c) [172.3, 188.3] d) [177.9, 182.7] e) [174.2, 179.1]

Oppgave 33
Hvordan tolker man et 99% konfidensintervall for forventet maxpuls ved alder lik 25 år?

a) Det er 99% sannsynlig at maxpulsen til en vilkårlig 25-åring ligger i dette intervallet.
b) Det er 99% sannsynlig at alderen til en vilkårlig person ligger i dette intervallet.
c) I 99 av 100 tilfeller vil et slikt intervall inneholde maxpulsen til en vilkårlig 25-åring.
d) Det er 99% sannsynlig at intervallet dekker den forventede maxpulsen til 25-åringer.
e) Det er 99% sannsynlig at den sanne effekten av alder på maxpuls ligger i dette intervallet.

Om befolkningen i Oslo (Oppgave 34-36)

Bystyret i Oslo ville sammenlikne aldersfordelingen av de som bor i tre bydeler i Oslo. I et utvalg på 1000 personer ble antall personer i de ulike aldersgruppene og bydelene talt opp. Tallene er gitt i tabellen nedenfor:

	Alder 0-19	**Alder 20-39**	**Alder 40-65**	**>66**	**Total**
Gamle Oslo	60	156	85	16	317
Grünerløkka	53	193	83	18	347
Nordstrand	84	87	118	47	336
Total	197	436	286	81	1000

Bystyret ville teste om aldersfordelingen er uavhengig av bydel i Oslo ved hjelp av en kji-kvadrat test. Nullhypotesen er at aldersfordeling og bydel er uavhengige.

Oppgave 34

Hvor mange frihetsgrader har testobservatoren Q for kji-kvadrattesten?

a) 12 b) 6 c) 10 d) 8 e) 4

Oppgave 35

Hva er estimert forventet antall innbyggere i bydelen Grünerløkka i aldersgruppen 20-39 år dersom nullhypotesen er sann?

a) 151.3 b) 193.1 c) 160.7 d) 436.0 e) 347.0

R Commander gav følgende tabeller over henholdsvis forventede antall og bidragene til testobservatoren Q fra de ulike kombinasjonene av bydel og aldersgruppe. (En av verdiene i den siste tabellen er erstattet med en stjerne)

```
# Expected Counts
              Alder 0-19    Alder 20-39   Alder 40-65  Alder >66
Gamle Oslo        62.449      138.212         90.662      25.677
Grünerløkka       68.359     (svar oppg 35)   99.242      28.107
Nordstrand        66.192      146.496         96.096      27.216

# Chi-square Components
               Alder 0-19    Alder 20-39  Alder 40-65  Alder >66
Gamle Oslo              *         2.29          0.35       3.65
Grünerløkka          3.45        11.50          2.66       3.63
Nordstrand           4.79        24.16          4.99      14.38
```

Oppgave 36

Hva er bidraget til testobservatoren Q fra bydelen Gamle Oslo i aldersgruppen 0-19 år (Tallet er erstattet med en stjerne i den siste tabellen ovenfor).

a) 40.54 b) 30.23 c) 2.60 d) 1.12 e) 0.10

Om vannkvalitet (Oppgave 37-40)
Vi ønsker å undersøke hvordan konsentrasjonen (Y) av E.coli bakterier avhenger av dybden (X) i en innsjø. Vannprøver ble tatt fra overflaten og fra 1 til 8 meters dybde. I hver prøve ble konsentrasjonen av E.coli bakterier målt. En lineær modell $y_i = \alpha + \beta x_i + \epsilon_i$ der $\epsilon_i \sim N(0, \sigma)$ ble antatt og resultatene fra en regresjonsanalyse i R Commander er gitt nedenfor.

```
Coefficients:
             Estimate Std. Error t value
(Intercept) 202.3778     *         *
Dybde        -1.9833     *         *

s: 2.077 on 7 degrees of freedom
Multiple R-squared: 0.8865,
Adjusted R-squared: 0.8703
F-statistic:  54.7 on 1 and 7 DF,  p-value: 0.00015
```

Oppgave 37
Hvordan tolker vi at $\hat{\beta} = -1.98$?

a) For hver meter man kommer nærmere overflaten så er den estimerte, forventede økningen i E.coli-konsentrasjonen lik 1.98.
b) Når E.coli-konsentrasjonen øker med en enhet vil dybden avta med ca 1.98 meter.
c) E.coli-konsentrasjonen ved dybde 0 (dvs i overflaten) forventes å være tilnærmet 1.98.
d) Når dybden øker med én meter er den estimerte økningen i forventet E-coli-konsentrasjon lik 1.98.
e) Reduksjonen i antall meters dybde som forventes dersom man reduserer E.coli-konsentrasjonen med 1 enhet er estimert til 1.98.

Oppgave 38
Anta at du vil teste hypotesen $H_0 : \beta = 0$ mot $H_1 : \beta \neq 0$ med et testnivå lik 1%. Hvilken tabellverdi fra t-tabellen må du da sammenlikne med?

a) 2.998 b) 2.821 c) 3.499 d) 3.250 e) 3.355

Oppgave 39
Minste kvadraters estimatoren for β i en lineærmodell av typen $y_i = \alpha + \beta x_i + \epsilon_i$ er såkalt forventningsrett. Hva legger vi i begrepet forventningsrett?

a) Når vi estimerer β med denne estimatoren får vi alltid det vi forventer å få.
b) Denne estimatoren vi i det lange løp i gjennomsnitt gi riktig verdi β.
c) Forventningen til β er lik estimatoren.
d) Estimatoren har en skjevhet, men den er veldig liten.
e) Variansen til estimatoren er liten.

Oppgave 40
For en viss dybde ble det beregnet et 95% konfidensintervall for forventet E.coli-konsentrasjon. Intervallet ble [195.0, 207.8]. Hva blir et 95% prediksjonsintervall for den samme dybden?

a) [193.3, 209.5] b) [194.2, 208.6] c) [191.2, 211.6] d) [198.0, 204.8] d) [181.3, 221.5]

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					