Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 17. december 2017

Kursus navn og nr: **Introduktion til Statistik (02402)**

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

_____ _____ _____

(studienummer) (underskrift) (bord nr)

Opgavesættet består af 30 spørgsmål af "multiple choice" typen fordelt på 18 opgaver. Besvarelserne af "multiple choice"spørgsmålene anføres i det i CampusNet uploadede svarark (på 6 seperate sider), med numrene på de svarmuligheder, du mener er de korrekte.

Der gives 5 point for et korrekt "multiple choice" svar og -1 for et ukorrekt svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller andet type svar angives, tæller det ikke med i besvarelsen. Envidere, hvis mere end et svar angives, hvilket faktisk er teknisk muligt i online-systemet, så tæller det ikke med (dvs. giver "0 point"). Det antal point, der kræves for, at et sæt anses for tilfredstillende besvaret, afgøres endeligt ved censureringen.

Den endelige besvarelse af opgaverne gøres ved at udfylde og online-aflevere svararket via CampusNet. Skemaet her er KUN et nød-alternativ til dette (husk at angive dit studienummer på din besvarelse, hvis du afleverer skemaet).

Opgave	I.1	II.1	II.2	III.1	III.2	IV.1	IV.2	V.1	V.2	V.3
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

Opgave	VI.1	VI.2	VII.1	VIII.1	VIII.2	IX.1	IX.2	IX.3	X.1	XI.1
Spørgsmål	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Svar										

Opgave	XII.1	XIII.1	XIII.2	XIV.1	XV.1	XVI.1	XVI.2	XVII.1	XVII.2	XVIII.1
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Sættet består af 26 sider.

Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at ideen med opgaverne er, at der er ét og kun ét rigtigt svar på de enkelte spørgsmål. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde.*

Opgave I

Vi betragter sammenhørende målinger af 2 stokastiske variable, X og Y, der begge kan antages normalfordelte.

x	38	35	47	38	42	41	48	35
y	25	21	26	23	28	27	29	18

Data kan indlæses i R ved:

```
x <- c(38, 35, 47, 38, 42, 41, 48, 35)
y <- c(25, 21, 26, 23, 28, 27, 29, 18)
```

Spørgsmål I.1 (1)

Et estimat af korrelationen, ρ, mellem X og Y bliver:

1 ☐ $\hat{\rho} = 0.12$

2 ☐ $\hat{\rho} = 0.22$

3 ☐ $\hat{\rho} = 0.64$

4 ☐ $\hat{\rho} = 0.73$

5 ☐ $\hat{\rho} = 0.82$

Fortsæt på side 3

Opgave II

En biolog evaluerer 3 forskellige diæters betydning for vægtændring i mus. I forsøget inkluderer man også 4 forskellige typer genetisk modificerede mus (strains), da dette kan have betydning for vægtændringen. Således indgår 4 forskellige typer mus for hver af de 3 diæter, dvs. i alt 12 mus. Man måler vægtændring efter 5 uger for hver diæt, benævnt Y_{ij} (i gram). Da vægtændringen kan antages normalfordelt, vælger man at analysere data ud fra følgende model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}.$$

I modellen angiver α_i effekten af diæt i ($i = 1, 2, 3$) og β_j ($j = 1, 2, 3, 4$) angiver effekt af type mus. Endelig er μ gennemsnittet og ε_{ij} er modellens afvigelser, der antages uafhængige og normalfordelt med middelværdi 0 og standardafvigelse σ_ε.

Spørgsmål II.1 (2)

Angiv den kritiske værdi, når man ønsker at teste om de 3 forskellige diæters betydning for vægtændring i middel kan antages ens når signifikansniveau $\alpha = 0.05$ anvendes.

1 ☐ 12.20

2 ☐ 5.14

3 ☐ 1.96

4 ☐ 3.81

5 ☐ 4.35

Spørgsmål II.2 (3)

Man har nu estimeret modellens parametre ved brug af R, og finder at både diæt og type mus er statistisk signifikante. Antag at man har gemt modellens residualer, ε_{ij}, i en vektor benævnt `resi`, og bruger R til at analysere disse. Hvilken af nedenstående påstande er ikke korrekt?

1 ☐ Kommandoen `qqnorm(resi)` afbilder normal scores for ε_{ij}

2 ☐ Kommandoen `plot(ecdf(resi))` afbilder den kumulative fordeling for ε_{ij}

3 ☐ Kommandoen `sum(resi*resi)/6` giver et estimat af variansen for ε_{ij}

4 ☐ Kommandoen `qnorm(resi)` giver et normalfordelingstest af ε_{ij}

5 ☐ Kommandoen `sum(resi)/length(resi)` giver estimat af middelværdien for ε_{ij}

Fortsæt på side 4

Opgave III

I et studie blev 605 forsøgspersoner, der tidligere havde haft hjertesygdom, randomiseret til en af 2 mulige diæter, som man ønskede at belyse effekten af. Efter en 4 års observationsperiode blev forsøgspersonerne klassificeret efter helbredsstatus (I) død, (II) kræft, (III) anden sygdom, eller (IV) rask.

Helbredsstatus

	I	II	III	IV	I alt
Diæt A	15	24	25	239	303
Diæt B	7	14	8	273	302
I alt	22	38	33	512	605

Nulhypotesen i studiet var, at der ingen sammenhæng er mellem diæt og helbredsstatus.

Spørgsmål III.1 (4)

Angiv hvordan den sædvanlige teststørrelse er fordelt, når vi antager at nulhypotesen er korrekt:

1 ☐ Den sædvanlige teststørrelse følger en χ^2-fordeling med 8 frihedsgrader

2 ☐ Den sædvanlige teststørrelse følger en F-fordeling med (1, 603) frihedsgrader

3 ☐ Den sædvanlige teststørrelse følger en t-fordeling med 4 frihedsgrader

4 ☐ Den sædvanlige teststørrelse følger en t-fordeling med 302 frihedsgrader

5 ☐ Den sædvanlige teststørrelse følger en χ^2-fordeling med 3 frihedsgrader

Spørgsmål III.2 (5)

Vi betragter nu andele af personer, der for hver diæt, er raske efter den 4-årige observationsperiode. Der ønskes et 95% konfidensinterval for forskel mellem andele, der er raske, på de 2 diæter. Angiv hvilket af følgende kald i R der giver dette:

1 ☐ `prop.test(x=c(512), n=c(605), correct=FALSE)`

2 ☐ `prop.test(x=c(303,302), n=c(512,512), correct=FALSE)`

3 ☐ `prop.test(x=c(239,273), n=c(303,302), correct=FALSE)`

4 ☐ `prop.test(x=c(239,273), n=c(605,605), correct=FALSE)`

5 ☐ `prop.test(x=c(239,273), n=c(512,512), correct=FALSE)`

Fortsæt på side 5

Opgave IV

I produktion af en fødevare indgår 3 delprocesser A, B og C. Den tid (i timer), det tager at udføre hver delproces, er stokastiske variable og benævnes X_A, X_B og X_C. Det antages, at X_A, X_B og X_C alle er normalfordelte og givet ved $X_A \sim N(12, 2^2)$, $X_B \sim N(25, 3^2)$ og $X_C \sim N(42, 4^2)$. Desuden antages, at tiden i hver delproces er uafhængig af de andre tider.

Den samlede produktionstid, Y, er defineret ved

$$Y = X_A + X_B + X_C.$$

Spørgsmål IV.1 (6)

Beregn nu sandsynligheden for at den samlede produktionstid, Y, overstiger 85 timer:

1 ☐ 0.0081

2 ☐ 0.1080

3 ☐ 0.1326

4 ☐ 0.4180

5 ☐ 0.6301

Spørgsmål IV.2 (7)

En ingeniør formår nu at optimere processen, således at den forbedrede produktionstid, Y^*, bliver

$$Y^* = 0.9 \cdot X_A + 0.8 \cdot X_B + X_C,$$

hvor X_A, X_B og X_C er defineret som i forrige spørgsmål.

Angiv nu variansen for Y^*:

1 ☐ $\mathrm{V}(Y^*) = (0.9 + 0.8 + 1) \cdot (2^2 + 3^2 + 4^2)$

2 ☐ $\mathrm{V}(Y^*) = (0.9^2 + 0.8^2 + 1^2) \cdot (2^2 + 3^2 + 4^2)$

3 ☐ $\mathrm{V}(Y^*) = 0.9 \cdot 2^2 + 0.8 \cdot 3^2 + 1 \cdot 4^2$

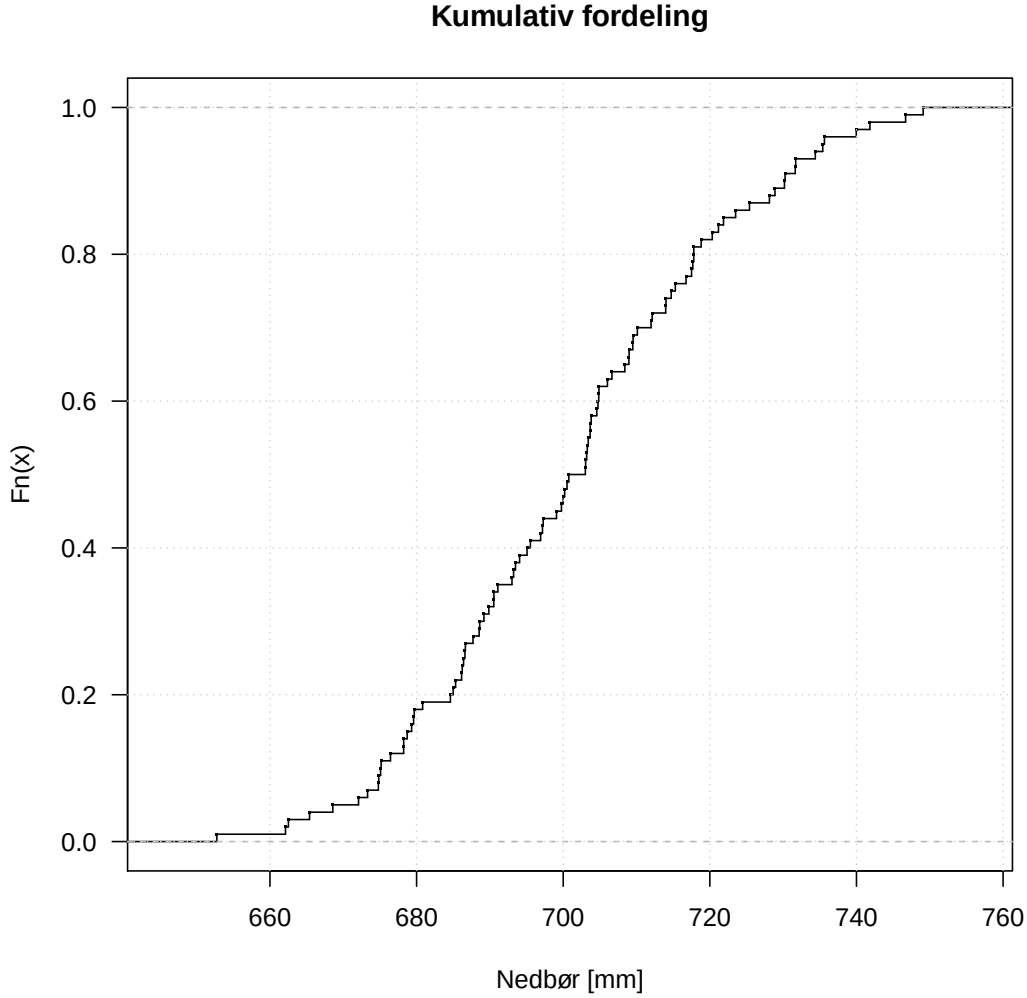
4 ☐ $\mathrm{V}(Y^*) = 0.9^2 \cdot 2^2 + 0.8^2 \cdot 3^2 + 1^2 \cdot 4^2$

5 ☐ $\mathrm{V}(Y^*) = 2^2 + 3^2 + 4^2$

Fortsæt på side 6

Opgave V

I et område har man registreret mængde nedbør (i mm) hvert år de sidste 100 år. Man kan antage, at den årlige nedbørsmængde er uafhængig fra år til år. Den kumulative fordeling for den årlige nedbørsmængde er vist i nedenstående figur.



Desuden oplyses følgende opsummering af data i R, hvor de 100 årlige målinger er gemt i variablen `rainfall`:

```
> var(rainfall)
[1] 412.7042
> summary(rainfall)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  652.8   686.6   701.9   701.3   714.9   749.1
```

Fortsæt på side 7

Spørgsmål V.1 (8)

Hvilket af nedenstående udsagn er ikke korrekt?

1 ☐ Estimatet af standardafvigelsen på det estimerede gennemsnit, $\hat{\sigma}_{\bar{X}}$, bliver $\frac{\sqrt{412.7042}}{10}$ mm

2 ☐ 50% fraktilen for de 100 målinger er 701.9 mm

3 ☐ Stikprøvestandardafvigelsen, s, for de 100 målinger er $\sqrt{412.7042}$ mm

4 ☐ 50% af de 100 målinger ligger mellem 686.6 og 714.9 mm

5 ☐ Variationskoefficienten (coefficient of variation) for de 100 målinger estimeres til $\frac{412.7042}{701.9}$

Spørgsmål V.2 (9)

Efterfølgende vil man beregne et 95% konfidensinterval for variansen af nedbørsmængden baseret på de 100 målinger, som man antager normalfordelte. Angiv nu 95% konfidensinterval for variansen af nedbørsmængden:

1 ☐ $[\frac{20.31512^2 \cdot 134.6416}{99}; \frac{20.31512^2 \cdot 69.22989}{99}]$

2 ☐ $[\frac{20.31512^2 \cdot 99}{134.6416}; \frac{20.31512^2 \cdot 99}{69.22989}]$

3 ☐ $[\frac{412.7042 \cdot 99}{128.422}; \frac{412.7042 \cdot 99}{73.36108}]$

4 ☐ $[\frac{412.7042 \cdot 99}{123.2252}; \frac{412.7042 \cdot 99}{77.04633}]$

5 ☐ $[\frac{20.31512 \cdot 99}{123.2252}; \frac{20.31512 \cdot 99}{77.04633}]$

Fortsæt på side 8

Spørgsmål V.3 (10)

Vi fortsætter med problemstillingen fra forrige side. Der oplyses nu, at følgende R kode er blevet kørt:

```
k = 10^5
Q5 = function(x){ quantile(x, 0.95) }
samples = replicate(k, sample(rainfall, replace = TRUE))
simvalues = apply(samples, 2, Q5)
interval = quantile(simvalues, c(0.025,0.975))
```

Der giver resultatet:

```
> interval
    2.5%    97.5%
728.9515 742.0814
```

Hvad er blevet beregnet og gemt i vektoren `interval`?

1 ☐ Et 95% konfidensinterval for middelværdien af nedbørsmængden (parametrisk bootstrap)

2 ☐ Et 95% konfidensinterval for 5% fraktilen af nedbørsmængden (parametrisk bootstrap)

3 ☐ Et 95% konfidensinterval for 95% fraktilen af nedbørsmængden (ikke-parametrisk bootstrap)

4 ☐ Et 95% konfidensinterval for 2.5% og 97.5% fraktilen af nedbørsmængden (ikke-parametrisk bootstrap)

5 ☐ Et 95% konfidensinterval for 2.5% og 97.5% fraktilen af nedbørsmængden (parametrisk bootstrap)

Fortsæt på side 9

Opgave VI

Vi betragter et forsøg, der kan resultere i et af to mulige udfald, benævnt A eller B. Sandsynligheden for udfaldet A benævnes $P(A)$. Vi får per definition at sandsynligheden for udfaldet B er $P(B) = 1 - P(A)$.

Spørgsmål VI.1 (11)

Vi vil nu observere en stokastisk variabel, X, som tæller antallet af udfald med resultatet A ud af $n = 300$ uafhængige forsøg. Hvis vi antager at $P(A) = 0.40$ er det samme i hvert af de n forsøg, hvad bliver så det forventede antal, $\mathrm{E}(X)$, og varians, $\mathrm{V}(X)$?

1 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot (1 - 0.4)$ og $\mathrm{V}(X) = 300^2 \cdot 0.4$

2 ☐ $\mathrm{E}(X) = 300 \cdot 0.4$ og $\mathrm{V}(X) = 300^2 \cdot 0.4 \cdot 0.6$

3 ☐ $\mathrm{E}(X) = 300 \cdot 0.4$ og $\mathrm{V}(X) = 300 \cdot 0.4 \cdot 0.6$

4 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot 0.6$ og $\mathrm{V}(X) = 300^2 \cdot 0.4^2 \cdot 0.6^2$

5 ☐ $\mathrm{E}(X) = 300 \cdot 0.4 \cdot 0.6$ og $\mathrm{V}(X) = 300 \cdot 0.4^2 \cdot 0.6^2$

Spørgsmål VI.2 (12)

Uafhængigt af dit svar i forrige spørgsmål, vil vi nu estimere sandsynligheden $P(A)$ baseret på de $n = 300$ forsøg. I $n = 300$ uafhængige forsøg finder man at i 120 tilfælde er udfaldet A og i 180 tilfælde er udfaldet B. Man vil nu estimere et 95% konfidensinterval for sandsynligheden $P(A)$. Dette bliver:

1 ☐ [0.33, 0.48]

2 ☐ [0.35, 0.46]

3 ☐ [0.35, 0.42]

4 ☐ [0.31, 0.53]

5 ☐ [0.29, 0.54]

Fortsæt på side 10

Opgave VII

En ingeniør undersøger kvaliteten af et batch råvarer. Kvalitetskravet er, at renheden af råvaren skal være mindst 90%.

Ingeniøren tager nu en tilfældig stikprøve på 10 uafhængige prøver fra batchen og gemmer måleværdierne (i %) af renheden i en vektor `x`.

Han kører nu følgende kode i R

```
> x <- c(90.6, 90.3, 88.9, 87.5, 87.6, 88.1, 87.5, 88, 88, 89.6)
> n <- length(x)
> tobs <- (mean(x) - 90) / (sd(x) / sqrt(n))
> pt(tobs, df=n-1)
```

Hvilket giver følgende output:

```
[1] 0.002279236
```

Spørgsmål VII.1 (13)

Baseret på ovenstående beregninger, samt en antagelse om at råvarens renhed er normalfordelt og anvendelse af signifikansniveau $\alpha = 0.05$, hvad kan ingeniøren så konkludere?

1 ☐ Ingeniøren kan konkludere, at renheden af råvaren i middel er mindst 88.6%.

2 ☐ Ingeniøren kan konkludere, at renheden af råvaren i middel højst er 88.6%.

3 ☐ Ingeniøren har med 99.7% sandsynlighed vist, at renheden af råvaren i middel er 90%.

4 ☐ Ingeniøren kan antage, at renheden af råvaren i middel er 90%.

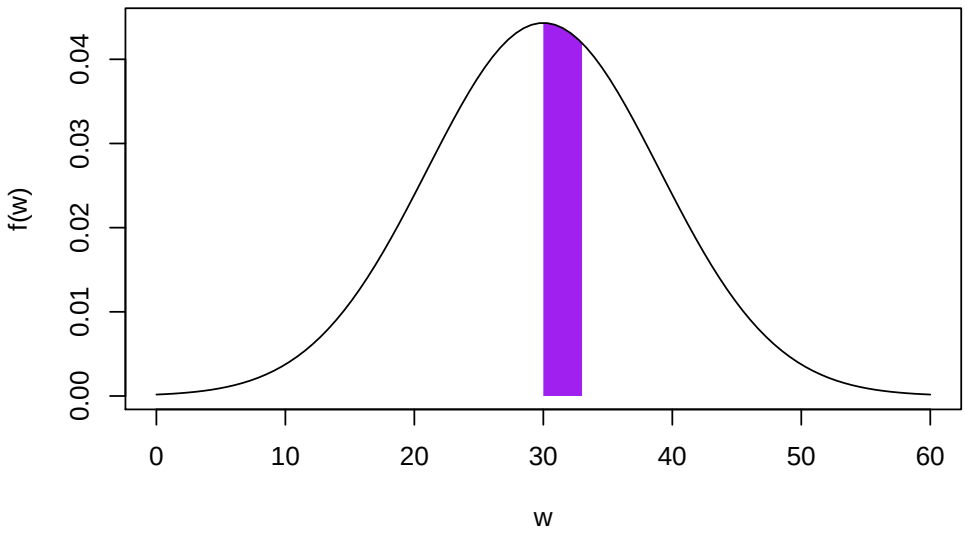
5 ☐ Ingeniøren kan afvise, at renheden af råvaren i middel er 90%.

Fortsæt på side 11

Opgave VIII

Vi betragter en stokastisk variabel W, der har tæthedsfunktion $f(w) = \frac{1}{9\sqrt{2\pi}} e^{-\frac{(w-30)^2}{162}}$.

Tæthedsfunktionen er afbildet i nedenstående figur, hvor også sandsynligheden $P(30 < W < 33)$ er markeret.



Spørgsmål VIII.1 (14)

Beregn nu sandsynligheden $P(30 < W < 33)$:

1 ☐ 0.09

2 ☐ 0.13

3 ☐ 0.24

4 ☐ 0.34

5 ☐ 0.84

Fortsæt på side 12

Spørgsmål VIII.2 (15)

Vi forestiller os nu, at der tages 3 stikprøver (benævnt A, B og C), hvor alle de tre stikprøver er taget fra fordelingen karakteriseret ved tætheden $f(w) = \frac{1}{9\sqrt{2\pi}} e^{-\frac{(w-30)^2}{162}}$ som i forrige spørgsmål.

Stikprøve A består af $n_A = 10$ målinger og estimatet af middelværdien benævnes $\hat{\mu}_A$. Stikprøve B består af $n_B = 30$ målinger og estimatet af middelværdien benævnes $\hat{\mu}_B$. Stikprøve C består af $n_C = 100$ målinger og estimatet af middelværdien benævnes $\hat{\mu}_C$.

Man interesserer sig især for, om stikprøvegennemsnittet kan overstige værdien 33, når den sande middelværdi i populationen er 30.

Hvilket af følgende udsagn er korrekt?

1 ☐ $P(\hat{\mu}_A \geq 33) > P(\hat{\mu}_B \geq 33)$

2 ☐ $P(\hat{\mu}_C \geq 33) > P(\hat{\mu}_A \geq 33)$

3 ☐ $P(\hat{\mu}_C \geq 33) = P(\hat{\mu}_B \geq 33)$

4 ☐ $P(\hat{\mu}_A \geq 33) = P(\hat{\mu}_B \geq 33) \cdot P(\hat{\mu}_C \geq 33)$

5 ☐ $P(\hat{\mu}_A \geq 33) = \frac{1}{2} P(\hat{\mu}_B \geq 33)$

Fortsæt på side 13

Opgave IX

Udbyttet af en kemisk proces, Y_i, tænkes at afhænge lineært af reaktionstemperaturen, t_i. For at opnå større indsigt i sammenhængen, har man udført et forsøg og målt $n = 50$ sammenhørende værdier af Y_i og t_i. Man tænker sig, at følgende model kan give en rimelig beskrivelse af sammenhængen

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \varepsilon_i.$$

Afvigelserne i ovenstående model antages i.i.d. og normalfordelt med konstant varians, altså $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Relevant output fra analysen af data er givet:

```
Call:
lm(formula = y ~ t)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0816 -1.4994 -0.2493  1.5175  4.8506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.4919     2.7757  23.595   <2e-16 ***
t             0.1637     0.1103   1.485    0.144
---


Residual standard error: 2.296 on 48 degrees of freedom
Multiple R-squared:  0.04392,Adjusted R-squared:  0.024
F-statistic: 2.205 on 1 and 48 DF,  p-value: 0.1441
```

Spørgsmål IX.1 (16)

Hvilket af følgende udsagn er korrekt, når signifikansniveau $\alpha = 0.05$ anvendes?

1 ☐ Udbyttet stiger med 16.37% når temperaturen stiger med en grad

2 ☐ Der kan ikke påvises en signifikant lineær sammenhæng mellem temperatur og udbytte

3 ☐ Teststørrelsen for ingen effekt af temperatur på udbyttet (dvs. nulhypotesen $H_0 : \beta_1 = 0$) er 23.595

4 ☐ Et 95% konfidensinterval for temperaturens effekt, β_1, er [-0.132027, 0.4594821]

5 ☐ Korrelationen mellem temperatur og udbytte er 0.04392

Fortsæt på side 14

Spørgsmål IX.2 (17)

Vi fortsætter med opgaven fra forrige side. Det viser sig, at processens pH kan have indflydelse på udbyttet, og da man har målt den i forsøget, vælger man derfor at inddrage den i modellen, der bliver:

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot pH_i + \varepsilon_i.$$

Estimation af modellens parametre giver følgende output i R:

```
Call:
lm(formula = y ~ t + pH)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7253 -1.2818 -0.2978  1.0724  4.4488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.46756    4.09799  12.071 5.25e-16 ***
t            0.24113    0.09315   2.589   0.0128 *
pH           2.37090    0.50097   4.733 2.06e-05 ***
---
Signif. codes:  0  '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.91 on 47 degrees of freedom
Multiple R-squared:  0.3525,Adjusted R-squared:  0.3249
F-statistic: 12.79 on 2 and 47 DF,  p-value: 3.667e-05
```

Angiv estimater for modellens parametre, dvs. β_0, β_1, β_2 og σ_ε^2

1 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (4.09799,\ 0.24113,\ 2.37090,\ 0.3525)$

2 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (49.46756,\ 0.24113,\ 2.37090,\ 1.91^2)$

3 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (49.46756,\ 0.24113,\ 2.37090,\ 1.91 \cdot 47)$

4 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (4.09799,\ 0.09315,\ 0.50097,\ 1.91 \cdot 47)$

5 ☐ $(\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\sigma}_\varepsilon^2) = (2.37090,\ 0.50097,\ 4.733\ ,\ 1.91)$

Fortsæt på side 15

Spørgsmål IX.3 (18)

Vi fortsætter med problemstillingen fra forrige spørgsmål og modellen

$$Y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot pH_i + \varepsilon_i.$$

Angiv et 95% konfidensinterval for den effekt man får på udbyttet, når pH stiger én enhed:

1 ☐ $0.24113 \pm 2.01174 \cdot 0.09315$

2 ☐ $2.37090 \pm 2.01174 \cdot 0.50097$

3 ☐ $(49.46756 + 0.24113 + 2.37090) \pm 2.01174 \cdot (4.09799 + 0.09315 + 0.50097)$

4 ☐ $2.37090 \pm 0.509920 \cdot 0.50097$

5 ☐ $(49.46756 + 0.24113 + 2.37090) \pm 0.509920 \cdot 0.50097$

Opgave X

Antag, at der findes en terning med 10 sider, og hvor sandsynligheden for hvert af de 10 udfald, $1, 2, \ldots, 10$, er den samme. Betragt den diskrete stokastiske variabel X der har tæthedsfunktion $f(x) = 0.1$ for $x \in (1, 2, \ldots, 10)$.

Spørgsmål X.1 (19)

Angiv nu middelværdien af X:

1 ☐ $\frac{1}{(10-1)} \sum_{i=1}^{10} x_i = 6.11$

2 ☐ $\frac{1}{(10-6.11)} \sum_{i=1}^{10} |x_i - 6.11| = 6.48$

3 ☐ $\frac{1}{(10)} \sum_{i=1}^{10} (x_i - 6.11)^2 = 8.62$

4 ☐ $\sum_{i=1}^{10} \frac{10-1}{10} x_i \cdot 0.1 = 4.95$

5 ☐ $\sum_{i=1}^{10} x_i \cdot 0.1 = 5.50$

Fortsæt på side 16

Opgave XI

Udbyttet af en proces er $\mu = 60$ mg/l. Man vil udføre nogle specifikke ændringer i processen og vil gerne kunne påvise en effekt på middeludbyttet, såfremt dette ændres med mindst 5 mg/l (dvs. et tosidet test).

En ingeniør skal nu dimensionere en undersøgelse, hvor man belyser effekten af ændringerne. Man vil nu beslutte hvor stor stikprøve, der skal tages fra den ændrede process. Forsøget skal være stort nok til at kunne opdage den relevante effekt (5 mg/l) med styrke (eng: power) på 0.8 når man anvender signifikansniveau $\alpha = 0.05$. Det kan antages, at standardafvigelsen for en måling af udbyttet er $\sigma = 10$ mg/l.

Spørgsmål XI.1 (20)

Baseret på ovenstående oplysninger, og ved brug af funktionen `power.t.test` i R, kommer man frem til at der mindst skal tages n målinger, hvor n bliver:

1 ☐ $n \simeq 256$ målinger

2 ☐ $n \simeq 128$ målinger

3 ☐ $n \simeq 64$ målinger

4 ☐ $n \simeq 34$ målinger

5 ☐ $n \simeq 27$ målinger

Fortsæt på side 17

Opgave XII

I et studie undersøger man en eventuel kolesterolsænkende effekt af et produkt. 9 forsøgspersoner fik målt deres kolesterolniveau (benævnt `x1`). Efter 3 måneder ved brug af produktet, fik de samme 9 forsøgspersoner målt deres kolesterolniveau igen, (benævnt `x2`). Data er vist i nedenstående tabel:

Person	1	2	3	4	5	6	7	8	9
x1	63.5	66.7	59.2	57.4	63.9	63.2	60.7	62.6	63.3
x2	51.3	51.9	57.8	50.2	54.6	43.3	51.2	40.4	52.2

Man kører nu følgende kode i R. Man ønsker at teste, om ændringen over tid kan antages at være lig nul ($H_0 : \delta = 0$):

```
x1 <- c(63.5, 66.7, 59.2, 57.4, 63.9, 63.2, 60.7, 62.6, 63.3)
x2 <- c(51.3, 51.9, 57.8, 50.2, 54.6, 43.3, 51.2, 40.4, 52.2)
```

Output fra den sædvanlige analyse er givet nedenfor. Bemærk dog, at nogle af tallene i det sædvanlige output er erstattet af symboler A, B, C.

```
t = -5.6354, df = A, p-value = B
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.847799  C
sample estimates:
mean of the differences
-11.95556
```

Spørgsmål XII.1 (21)

Idet man anvender signifikansniveau $\alpha = 0.05$ kommer man til følgende konklusion:

1 ☐ Der kan påvises en effekt, idet $\mu_D = -11.95556$

2 ☐ Der kan ikke påvises en effekt, øvre grænse i konfidensintervallet er 7.063312

3 ☐ Der kan ikke påvises en effekt, nedre grænse i konfidensintervallet er -7.063312

4 ☐ Der kan påvises en effekt, idet p-værdien er $4.897 \cdot 10^{-4}$

5 ☐ Der kan påvises en effekt, idet p-værdien er $2.394 \cdot 10^{-4}$

Fortsæt på side 18

Opgave XIII

En biolog er interesseret i at undersøge effekten af 4 forskellige vækshæmmere, benævnt V_1, V_2, V_3 og V_4. De 4 vækshæmmere tilsættes prøver fra den samme cellelinje og vækst efter en uge måles Y_{ij} (i antal celler per cm^2). Der udføres 8 gentagne og uafhængige målinger for hver cellelinje, dvs. i alt 32 målinger. Da målingerne kan antages normalfordelt, vælger man at analysere data ud fra følgende variansanalysemodel

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

I modellen angiver α_i effekten af vækshæmmer i ($i = 1, 2, 3, 4$), μ er gennemsnittet og ε_{ij} er modellens afvigelser, der antages uafhængige og normalfordelt med middelværdi 0 og standardafvigelse σ_ε.

En variansanalyse for ovenstående model er givet nedenfor, dog er output her ufuldstændigt, idet nogle tal er erstattet af symbolerne A, B og C.

```
Analysis of Variance Table

Response: growth
          Df Sum Sq Mean Sq F value    Pr(>F)
treatment  A 281.07       B       C 0.0001409 ***
Residuals 28 268.46   9.588
```

Spørgsmål XIII.1 (22)

Hvad bliver den sædvanlige teststørrelse (markeret ved `C`), når man ønsker at teste om der er forskel i middel på den effekt de 4 vækshæmmere har?

1 ☐ 9.77

2 ☐ 7.23

3 ☐ 2.95

4 ☐ 4.57

5 ☐ 16.11

Fortsæt på side 19

Spørgsmål XIII.2 (23)

Vi ønsker nu at beregne et post hoc 95% konfidensinterval for forskel i middelværdi på vækst-hæmmer V_1 og V_2, benævnt $I_{0.95}(V_1 - V_2)$. Fra forsøgene er forskellen i middel på V_1 og V_2 estimeret til 4.5. Beregn nu $I_{0.95}(V_1 - V_2)$:

1 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.048 \cdot \frac{9.588}{12} \cdot \sqrt{28}$

2 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.048 \cdot \sqrt{9.588} \cdot \sqrt{2/8}$

3 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.306 \cdot \frac{\sqrt{9.588}}{\sqrt{12}}$

4 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 2.306 \cdot 9.588^2 \cdot \sqrt{1/8}$

5 ☐ $I_{0.95}(V_1 - V_2) = 4.5 \pm 1.960 \cdot \frac{9.588}{\sqrt{8}}$

Opgave XIV

Vi betragter en kontinuert stokastisk variabel X, hvor den velkendte fordelingsfunktion $F(x)$ er givet ved $P(X \leq x) = 1 - e^{-x/2}$, hvor $x > 0$.

Spørgsmål XIV.1 (24)

Angiv nu middelværdien for X:

1 ☐ $\frac{1}{2}$

2 ☐ 1

3 ☐ 2

4 ☐ $\frac{3}{2}$

5 ☐ 4

Fortsæt på side 20

Opgave XV

En biolog undersøger biodiversiteten i et område, og har blandt andet målt antal forskellige planter per 10 m^2 forskellige steder i området. Hun har i alt 30 uafhængige målinger, y_i, og disse er gemt som en vektor i statistikprogrammet R og benævnt `Yobs`.

Spørgsmål XV.1 (25)

Biologen vil gerne beregne et 95% konfidensinterval for variationskoefficient (coefficient of variation) for biodiversiteten (antal forskellige planter per 10 m^2) ved brug af ikke-parametrisk bootstrap. Hvilket af nedenstående forslag (kode i R) er mest hensigtsmæssigt til denne beregning?

1 ☐
```
samples = replicate(10000,rnorm(30,mean(Yobs),sd(Yobs))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

2 ☐
```
samples = replicate(10000,sample(Yobs,replace=TRUE))
results = apply(samples,2,var)/apply(samples,2,sd)
quantile(results, c(0.025,0.975))
```

3 ☐
```
samples = replicate(10000,rnorm(30,mean(Yobs),sd(Yobs))
results = apply(samples,2,var)/apply(samples,2,median)
quantile(results, c(0.025,0.975))
```

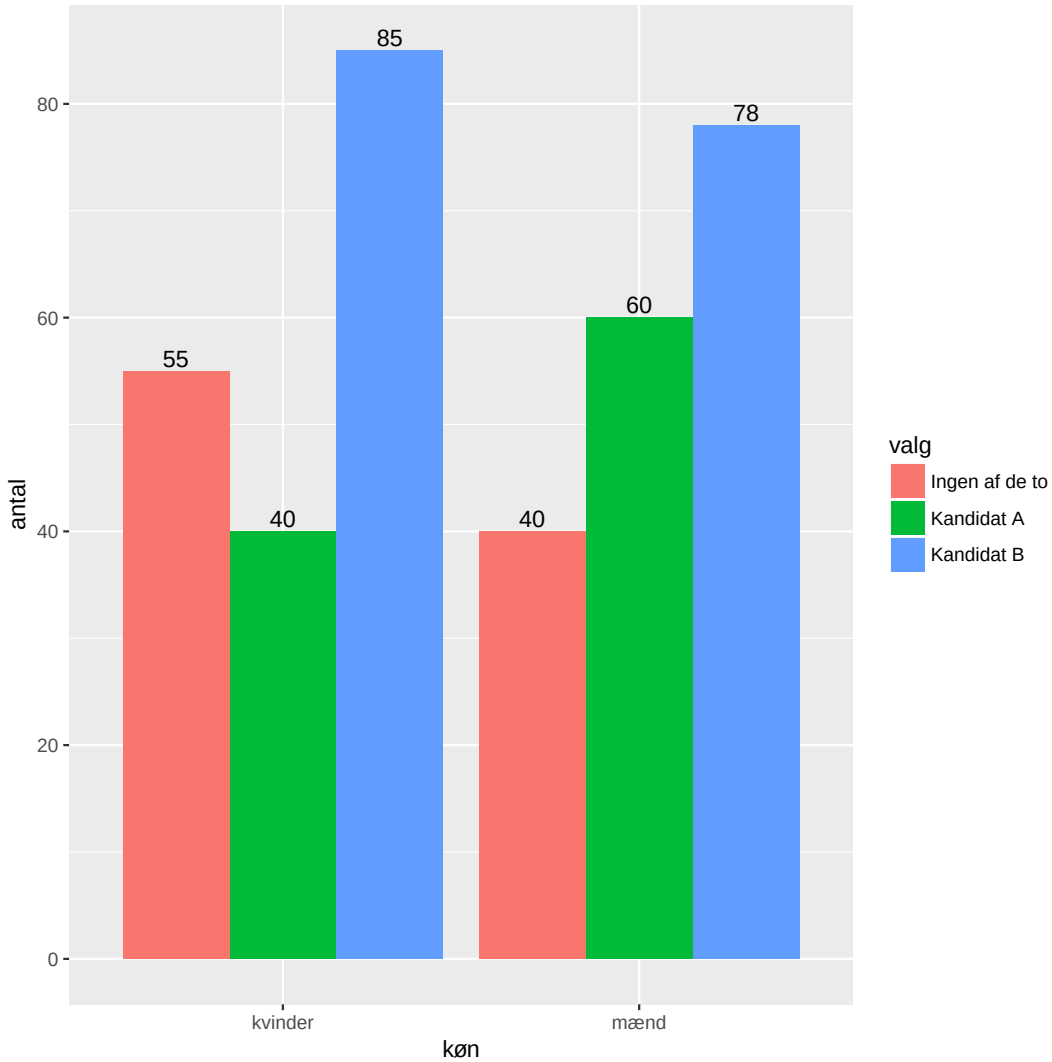
4 ☐
```
samples = replicate(10000,sample(Yobs,replace=FALSE))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

5 ☐
```
samples = replicate(10000,sample(Yobs,replace=TRUE))
results = apply(samples,2,sd)/apply(samples,2,mean)
quantile(results, c(0.025,0.975))
```

Fortsæt på side 21

Opgave XVI

I en undersøgelse blev 180 kvinder og 178 mænd bedt om at vurdere hvilke af 2 politiske kandidater, A eller B, de foretrak. Alternativt kunne man svare "ingen af de to". Fordelingen er vist i figuren herunder.



Fortsæt på side 22

Spørgsmål XVI.1 (26)

Det fremgår eksempelvis af figuren, at man observerer at 85 ud af de 180 deltagende kvinder foretrækker kandidat B. Såfremt kvinder og mænds svarfordeling er ens, hvor mange kvinder ud af de 180 ville man have forventet ville foretrække kandidat B?

1 ☐ $\frac{163}{358} \cdot \frac{95}{358} \cdot 358$

2 ☐ $\frac{100}{358} \cdot \frac{223}{358} \cdot 358$

3 ☐ $\frac{95}{358} \cdot \frac{190}{358} \cdot 358$

4 ☐ $\frac{163}{358} \cdot \frac{180}{358} \cdot 358$

5 ☐ $\frac{95}{358} \cdot \frac{180}{358} \cdot 358$

Spørgsmål XVI.2 (27)

Hvad bliver den sædvanlige teststørrelse, når man ønsker at teste, om der i middel er forskel i svarfordelingen for mænd og kvinder?

1 ☐ $\chi^2_{\text{obs}} = 5.9915$

2 ☐ $\chi^2_{\text{obs}} = 6.6581$

3 ☐ $\chi^2_{\text{obs}} = 16.212$

4 ☐ $\chi^2_{\text{obs}} = 8.3836$

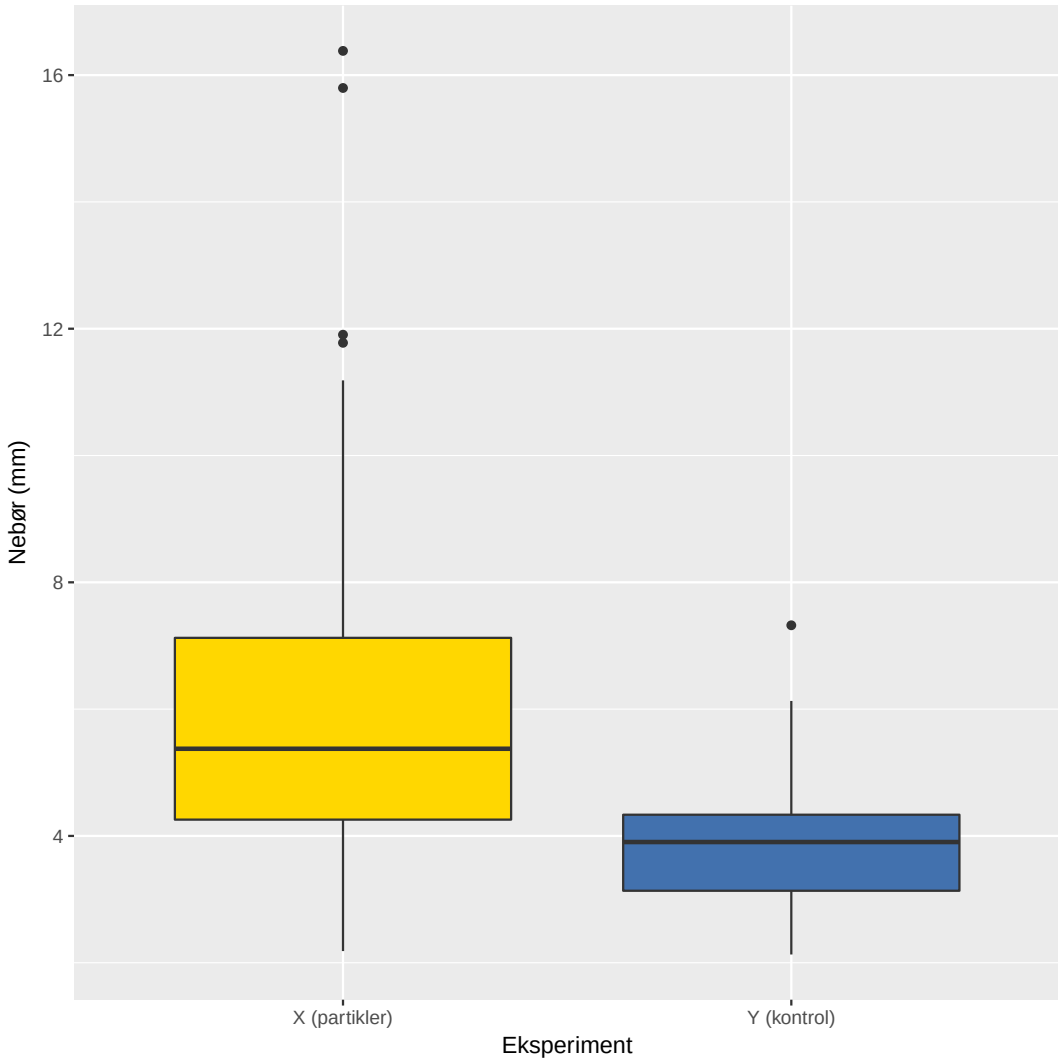
5 ☐ $\chi^2_{\text{obs}} = 4.5067$

Fortsæt på side 23

Opgave XVII

Skyskabning er en teknik, der kan anvendes for at øge nedbørsmængden. Skyerne skabes ved at sprøjte små partikler, så som aluminiumoxid, på skyer for at påvirke deres udvikling.

I et eksperiment ville man undersøge skyskabningens effekt ved brug af en ny type partikler. Man sammenlignede nedbørsmængde (mm nedbør per dag) for 35 dage hvor der var sprøjtet med partiklerne, benævnt X_i, $(i = 1, 2, \ldots, 35)$, med nedbørsmængde per dag på 30 dage, hvor der ikke var sprøjtet, benævnt Y_j, $(j = 1, 2, \ldots, 30)$. Der blev kun udført forsøg (målinger) på dage, hvor det var tilstrækkelig fugtighed i luften. Data fra forsøget er vist i nedenstående figur.



Fortsæt på side 24

Vi vil nu analysere data beskrevet på forrige side ved brug af R. Idet data x_i er gemt i vektoren `x` og data y_j er gemt i vektoren `y`, har man kørt følgende kode:

```
k <- 10^4
resultX <- replicate(k, sample(x, replace = TRUE))
resultY <- replicate(k, sample(y, replace = TRUE))
result <- apply(resultX, 2, median) - apply(resultY, 2, median)
quantile(result, c(0.5, 0.025,0.975))
```

Der giver resultatet

```
      50%      2.5%     97.5%
1.6283069 0.2843492 2.4233546
```

Spørgsmål XVII.1 (28)

Hvis man anvender signifikansniveau $\alpha = 0.05$ kan man baseret på ovenstående resultater konkludere:

1 ☐ Medianen for X er signifikant højere end medianen for Y.

2 ☐ Medianen for X er 62.8% højere end medianen for Y.

3 ☐ Nedbør ved X er 28.4% til 142.3% højere end nedbør ved Y.

4 ☐ Middelnedbør kan anses for at være ens for de to metoder.

5 ☐ Medianen for Y er $[0.28; 2.42]$ højere end medianen for X.

Fortsæt på side 25

Spørgsmål XVII.2 (29)

I et andet forsøg med skyskabning, har man undersøgt effekt af en ny og anden type partikler. Også her sammenlignes nedbørsmængden i forhold til at man ikke sprøjter med partikler. Således er der tale om et nyt og tilsvarende forsøg blot med en ny type partikler. Her har man dog valgt at transformere data med den naturlige logaritme, og opnår derved at data i begge grupper kan antages at være normalfordelte. Data er opsummeret i nedenstående tabel (log mm nedbør).

	Partikler, X (log mm nedbør)	Kontrol, Y (log mm nedbør)
Estimat middelværdi	$\hat{\mu}_X = 1.573$	$\hat{\mu}_Y = 1.314$
Estimat varians	$\hat{\sigma}_X^2 = 0.333$	$\hat{\sigma}_Y^2 = 0.171$
Antal observationer	$n_X = 35$	$n_Y = 30$

Idet man ønsker at teste om der er forskel i middel på de 2 grupper, dvs.

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

Det oplyses, at den sædvanlige teststørrelse under nul-hypotesen bliver 2.0958 med 61.19 frihedsgrader. Angiv p-værdi og konklusion, idet signifikansniveauet $\alpha = 0.05$ anvendes:

1 ☐ p-værdi $\simeq 0.82$ dvs. H_0 accepteres

2 ☐ p-værdi $\simeq 0.41$ dvs. H_0 forkastes

3 ☐ p-værdi $\simeq 0.21$ dvs. H_0 accepteres

4 ☐ p-værdi $\simeq 0.10$ dvs. H_0 forkastes

5 ☐ p-værdi < 0.05 dvs. H_0 forkastes

Fortsæt på side 26

Opgave XVIII

På et julemarked har man en tombola. I en tromle ligger 24 kugler. På 4 af kuglerne er der et billede af en stjerne. På de øvrige 20 kugler er der et billede af en nisse. Lotteriet går nu ud på, at man skal trække 2 kugler (uden tilbagelægning) fra tromlen. Hvis begge kugler viser et billede af en stjerne, har man vundet en præmie.

Spørgsmål XVIII.1 (30)

Vi deltager i spillet en enkelt gang. Angiv sandsynligheden for at vinde en præmie

1 ☐ $\frac{80}{276}$

2 ☐ $\frac{56}{276}$

3 ☐ $\frac{40}{276}$

4 ☐ $\frac{16}{276}$

5 ☐ $\frac{6}{276}$

SÆTTET ER SLUT. God Juleferie!