NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4255 Anvendt statistikk**

Faglig kontakt under eksamen: Anna Marie Holand

Tlf: 951 38 038

Eksamensdato: August 2016

Eksamenstid (fra–til):

Hjelpemiddelkode/Tillatte hjelpemidler: C: Gult, stemplet A4-ark med dine egne håndskrevne notater, Tabeller og formler i statistikk (Tapir forlag/Fagbokforlaget). Bestemt kalkulator.

Annen informasjon:

- I utskrift fra MINITAB er komma brukt som desimalseparator.
- Signifikansnivå 5% skal brukes hvis ikke annet er spesifisert.
- Alle svar må begrunnes.

Målform/språk: bokmål

Antall sider: 10

Antall sider vedlegg: 0

Kontrollert av:

Dato Sign

Oppgave 1 Behandling for reduksjon av stress

Det ble utført et studie der tre ulike behandlinger mot reduksjon av stress ble sammenlignet. I studien ble 30 deltakere tilfeldig delt inn i tre forskjellige behandlingsgrupper, gruppe A ble gitt en mental behandling, gruppe B ble gitt behandling med fysisk trening, og gruppe C ble gitt en medisinsk behandling for å redusere stress. En poengskår mellom 1 og 5 ble gitt etter at behandlingen var ferdig. Poengskåren angir hvor effektiv behandlingen var for å redusere deltakernes stressnivå, der en høy poengskår indikerer høy effektivitet i å redusere stress.

Data og deskriptive mål for behandlingspoengskår for de 3 forskjellige gruppene er gitt i henholdsvis tabell 1 og tabell 2.

Gruppe A	2	2	3	4	4	5	3	4	4	4
Gruppe B	4	4	3	5	4	1	1	2	3	3
Gruppe C	1	2	2	2	3	2	3	1	3	1

Tabell 1: Poengskår for stressreduksjon for gruppene A, B, C i behandling for reduksjon av stress.

Treatment group	Sample size	Mean	Standard deviation
Group A	10	3.5	0.972
Group B	10	3.0	1.333
Group C	10	2.0	0.816
Total	30	2.83	

Tabell 2: Deskriptive mål for gruppene A, B, C i datasettet for behandling for reduksjon av stress.

Først var det av interesse å sammenligne gruppen som fikk mental behandling (gruppe A) og gruppen som fikk behandling med fysisk trening (gruppe B).

a) La S_A og S_B være standardavvikene for gruppe A og gruppe B og la disse være estimatorer for σ_A og σ_B.

Test hypotesen

$$H_0 : \sigma_A = \sigma_B \text{ vs. } H_1 : \sigma_A \neq \sigma_B$$

ved å beregne et 95% konfidensintervall for $\frac{\sigma_A}{\sigma_B}$.

Basert på disse dataene, kan vi konkludere med at poengskårene for reduksjon i stress er forskjellige for de to behandlingsgruppene (gruppe A og gruppe B)? Skriv ned nullhypotesen og den alternative hypotesen og utfør en t-test for to grupper basert på de deskriptive målene i tabell 2.

Bruk signifikansnivå $\alpha = 0.05$. Spesifiser hvilke antagelser du gjør. Hva er konlusjonen fra testen?

b) En Wilcoxon rank-sum (Mann–Whitney) test ble utført på dataene fra gruppe A og gruppe B for å teste om poengskårene for reduksjon i stress er forskjellige for de to behandlingsgruppene (gruppe A og gruppe B).

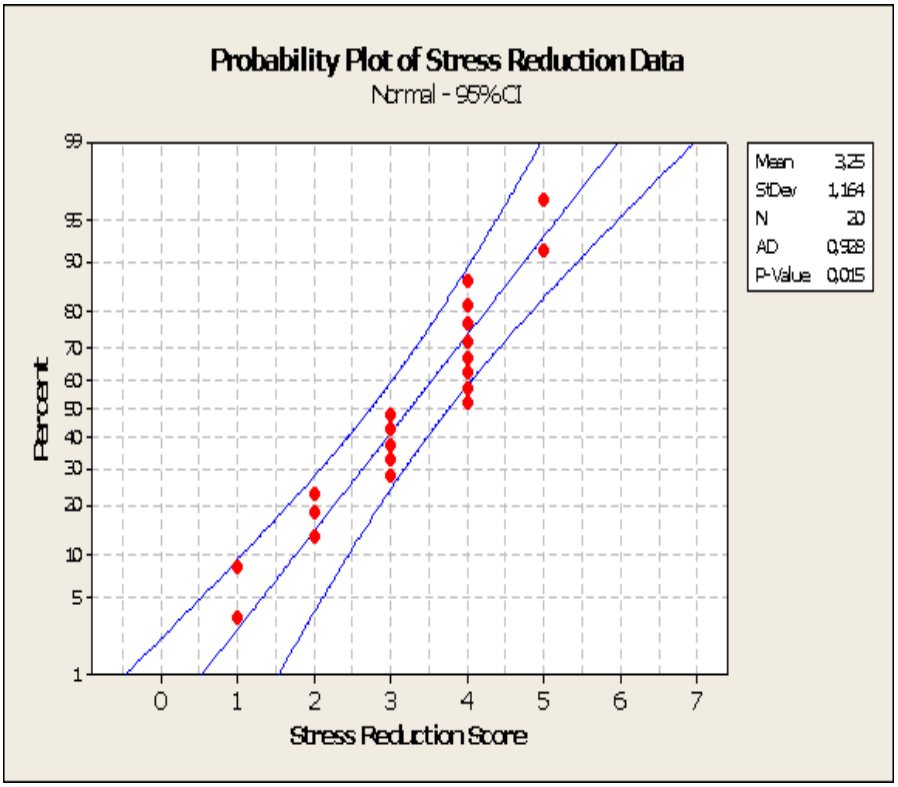
Skriv ned nullhypotesen og den alternative hypotesen for denne testen. Fra MINITAB-utskriften i figur 1, er testobservatoren beregnet til $W_1 = 116$. Vis hvordan denne verdien kan bli funnet fra dataene i tabell 1. Forklar kort hvordan verdien av W_1 blir brukt for å teste hypotesen.

Bruk signifikansnivå $\alpha = 0.05$.

Hvilke antagelser må du gjøre i denne testen?
Hviken av testene i a) (andre halvdel) og b) syns du er mest passende for dataene? Du kan basere svaret på figur 2.

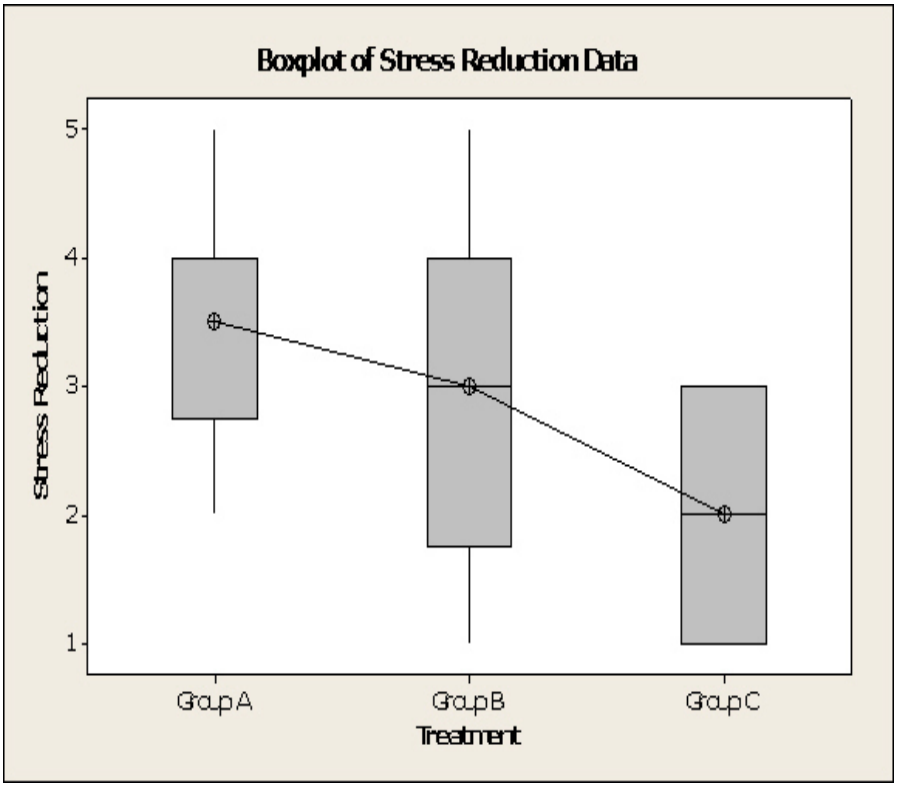
```
Mann-Whitney Test and CI

         N  Median
Group A  10   4,000
Group B  10   3,000

Point estimate for ETA1-ETA2 is 1
95,5 Percent CI for ETA1-ETA2 is (-1,000;1,999)
W = 116,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,4274
The test is significant at 0,4073 (adjusted for ties)
```

Figur 1: Utskrift fra en Wilcoxon rank-sum test (Mann-Whitney) for datasettet for behandling for reduksjon av stress.



Figur 2: Et normalplott basert på datasettet for behandling for reduksjon av stress, der dataene fra gruppe A og gruppe B er slått sammen.

I resten av oppgaven skal du bruke alle de tre stressbehandlingsgruppene, A, B og C. Et boksplott for de tre behandlingsgruppene er gitt i figur 3.



Figur 3: Boksplott for gruppene A, B og C i datasettet for behandling for reduksjon av stress.

c) Det er av interesse å finne ut om det er en forskjell i effekten av stressreduksjon mellom de tre behandlingsgruppene.

Beregn først kvadratsummene, SSA og SSE, ved å bruke de deskriptive målene i tabell 2.

Utfør en enveis variansanalyse for å teste nullhypotesen

$$H_0 : \mu_A = \mu_B = \mu_C.$$

Bruk signifikansnivå $\alpha = 0.05$. Hva er antagelsene du må gjøre i denne testen? Hva er konklusjonen av testen?

d) Fra resultatene i c) ble det besluttet å utforske videre om noen av forskjellene $\mu_A - \mu_B$, $\mu_A - \mu_C$ og $\mu_B - \mu_C$ er signifikant forskjellige fra 0.

Bruk Bonferroni-metoden for dette problemet. Family-wise error rate (FWER), dvs. sannsynligheten for å gjøre minst én type I feil, skal være kontrollert på nivå α. Bruk $\alpha = 0.05$.

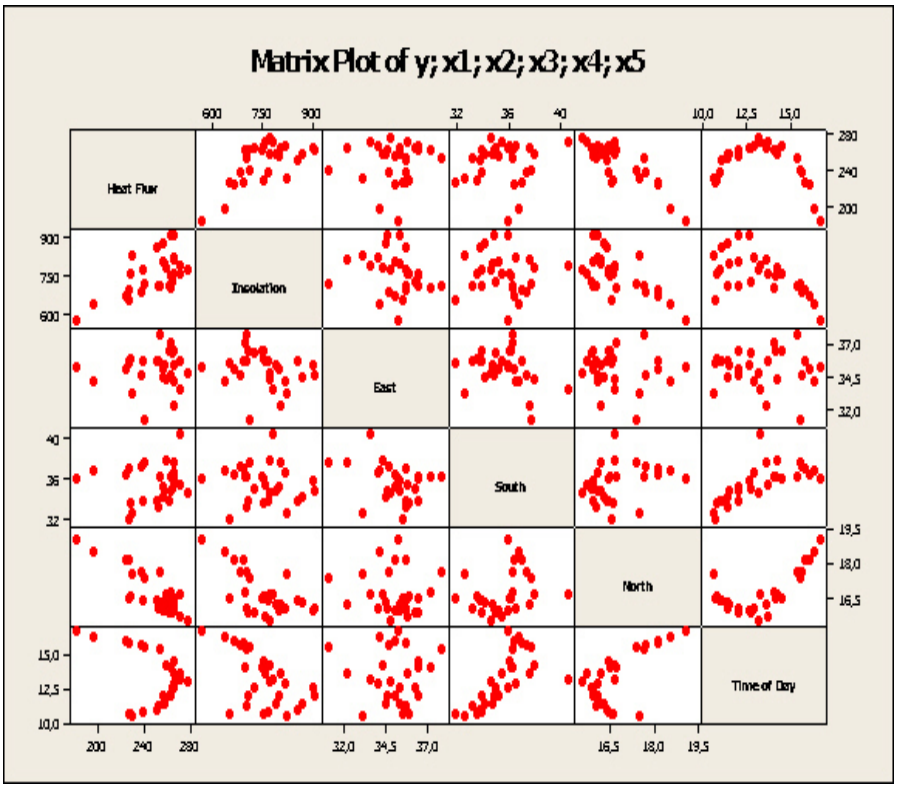
Oppgave 2 Varmegjennomstrømning

Ingeniører som jobber med energi er interessert i varmeenergi fra solen som del av utviklingen av solcelleenergi. Varmegjennomstrømning (heat flux) er raten varmeenergi som blir overført gjennom en gitt overflate per enhet tid. Varmegjennomstrømning blir målt som en del av en energitest. En ingeniør er interessert i å bestemme hvordan den totale varmegjennomstrømningen blir predikert av variablene: innstråling, posisjonene i øst, sør og nord for brennpunktene, og tiden på dagen. $n = 29$ varmeenergitester ble tatt.

Følgende beskrivelse er gitt.

- y: den totale varmegjennomstrømningen, i kilowatt
- x_1: innstrålingsverdien, i watt/m2
- x_2: posisjonen for brennpunktene i retning øst, i tommer
- x_3: posisjonen for brennpunktene i retning sør, i tommer
- x_4: posisjonen for brennpunktene i retning nord, i tommer
- x_5: tid på dagen dataene ble tatt

Et parvis spredningsplott og en korrelasjonsmatrise av variablene finnes i figur 4.



	y	x1	x2	x3	x4
x1	0,628				
x2	0,102	-0,204			
x3	0,112	-0,107	-0,329		
x4	-0,849	-0,634	-0,117	0,287	
x5	-0,351	-0,584	-0,065	0,697	0,685

Figur 4: Parvis spredningsplott (øvre del) og parvis Pearson korrelasjon (nedre del) mellom variablene y, x_1, x_2, x_3, x_4 og x_5 i varmegjennomstrømningsdatasettet.

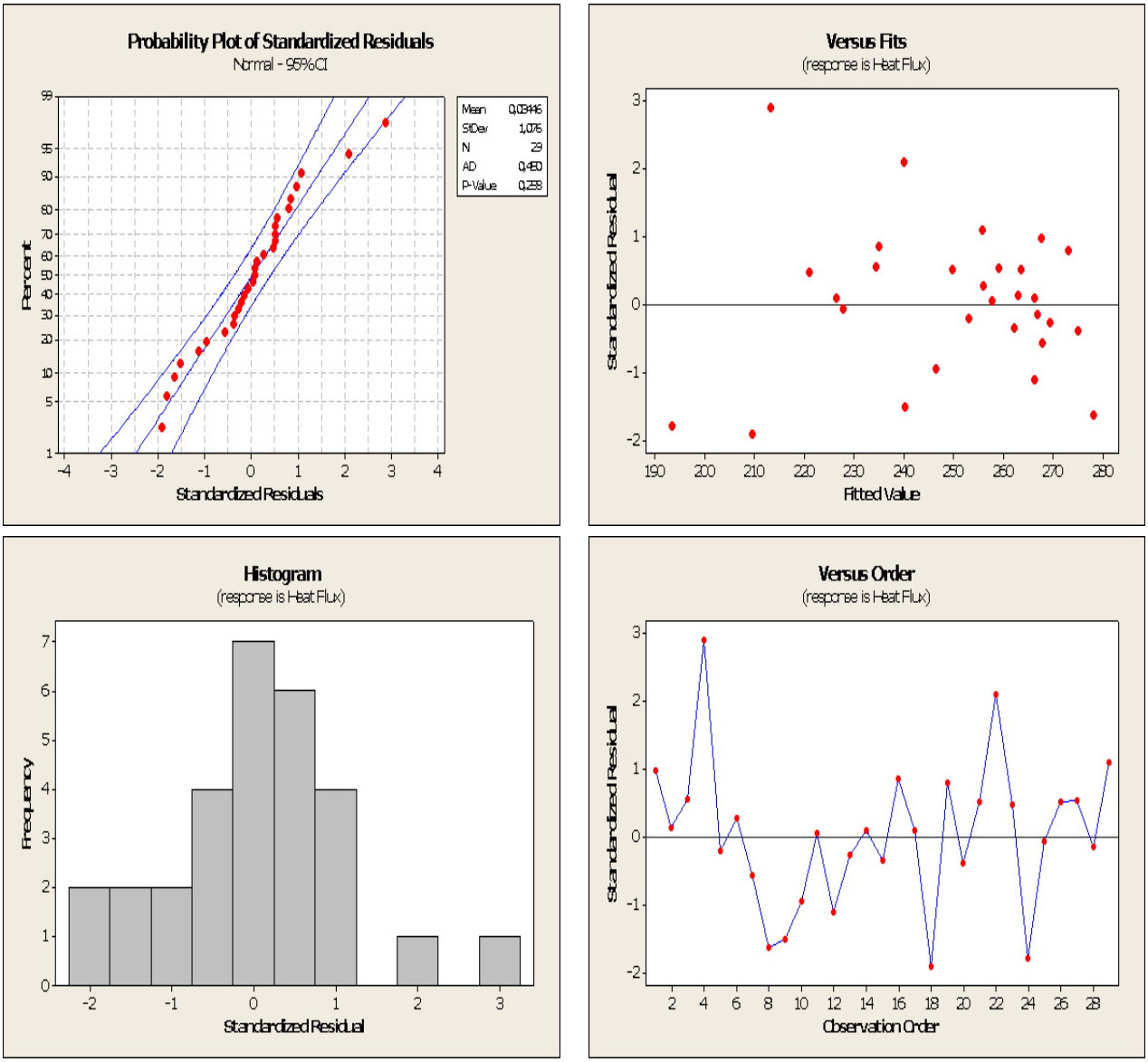
En multippel lineær regresjonsmodell ble tilpasset til dataene med y som respons og x_1, x_2, x_3, x_4 og x_5 som forklaringsvariabler. La (y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i} og x_{5i}) angi en observasjon fra varmeenergitest i, hvor $i = 1, ..., 29$. Definer den fulle modellen (modell A):

$$\text{Modell A} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \epsilon_i, \tag{1}$$

hvor ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, ..., n$. MINITAB-utskriften fra en statistisk analyse av modell A finnes i figur 5. Plott av standardiserte residualer finnes i figur 6.

```
Predictor       Coef  SE Coef      T      P
Constant      325,44    96,13   3,39  0,003
Insolation   0,06753  0,02899   2,33  0,029
East           2,552    1,248   2,04  0,053
South          3,800    1,461   2,60  0,016
North        -22,949    2,704  -8,49  0,000
Time of Day    2,417    1,808   1,34    ?

S = ?   R-Sq = ?%   R-Sq(adj) = 87,7%

PRESS = 3109,95   R-Sq(pred) = 78,82%

Analysis of Variance
Source          DF   SS       MS      F      P
Regression       5   ?        2639,1  40,84  0,000
Residual Error  23   ?        64,6
Total           28  14681,9
```

Figur 5: Utskrift fra tilpassing av den lineære regresjonsmodellen A i varmegjennomstrømningsdatasettet.



Figur 6: Residualplott (normal plott basert på standardiserte residualer i øvre venstre panel, standardiserte residualer mot tilpassede verdier i øvre høyre panel, histogram basert på standardiserte residualer i nedre venstre panel og standardiserte residualer mot rekkefølgen på observasjonene i nedre høyre panel) for regresjonsmodell A i varmegjennomstrømningsdatasettet.

a) Skriv ned den estimerte regresjonsligningen.

Nå skal vi se på den estimerte regresjonskoeffisienten for X_5, ***Tid på dagen***, i denne modellen. Er effekten av X_5, ***Tid på dagen***, signifikant i denne modellen?

Beregn R^2 og forklar hvordan du kan tolke denne verdien.

Hva er et passende estimat for σ? Gi den numeriske verdien av dette estimatet.

Er en signifikant del av variasjonen i dataene forklart av denne modellen? Skriv ned nullhypotesen og den alternative hypotesen. Velg en testobservator og utfør en hypotesetest. Bruk signifikansnivå $\alpha = 0.05$.

b) Hvilke modellantagelser ble gjort i model A in ligning (1)? Forklar og definer hva et residual er.

Figur 6 viser noen residualplott for model A. Indikerer residualplottene at disse antagelsene for lineær regresjon er oppfylt? Begrunn svaret ved å kommentere kort på plottene i figur 6. Hvordan ville de ideelt se ut hvis modellen er korrekt?

Vi er nå interessert i å sammenligne ulike regresjonsmodeller, der vi betrakter ulike kombinasjoner av forklaringsvariablene x_1, x_2, x_3, x_4 og x_5, i en "best subset" regresjon. Anta at et konstantledd, β_0, er med i regresjonsmodellen.

MINITAB-utskriften fra tilpasning av ulike modeller for dataene er presentert i figur 7. Hver rad i figur 7 svarer til en modell. Antallet forklaringsvariabler inkludert i hver modell (i tillegg til konstantleddet, β_0) finner du i kolonnen med navn $Vars$. De to beste modellene for hvert antall forklaringsvariabler er rapportert. X'ene indikerer hvilke variabler som er funnet i modellen.

				Mallows		x	x	x	x	x
Vars	R-Sq	R-Sq(adj)	R-Sq(pred)	Cp	S	1	2	3	4	5
1	72,1	71,0	66.9	38,5	12,328				X	
1	39,4	37,1	26.3	112,7	18,154	X				
2	85,9	84,8	81.4	9,1	8,9321			X	X	
2	82,0	80,6	74.2	17,8	10,076				X	X
3	87,4	85,9	79.0	7,6	8,5978		X	X	X	
3	86,5	84,9	81.4	9,7	8,9110	X		X	X	
4	89,1	87,3	80.6	5,8	8,1698	X	X	X	X	
4	88,0	86,0	79.3	8,2	8,5550	X		X	X	X
5	89,9	87,7	78.8	6,0	8,0390	X	X	X	X	X

Figur 7: Utskrift fra statistisk analyse av varmegjennomstrømningsdatasettet.

c) Basert på resultatet i figur 7, hvilken av disse regressjonsmodellene ville du valgt som den "beste" modellen for datasettet? Begrunn valget ditt.

For å svare på dette spørsmålet skal du også forklare hvordan R^2, R^2_{adj}, R^2_{pred}, Mallows C_p og S er definert og hvordan du kan bruke dem til å sammenligne de ulike regresjonsmodellene.

Oppgave 3 Tiden det tar å knyte skolisser

Et pilotstudie ble gjennomført der det ble målt tiden det tar, i sekunder, for barn på 9 år å knyte skolissene.

Totalt antall barn var $n = 250$. Tiden, gitt i intervaller, og antall barn i hvert intervall finnes i følgende tabell.

	1	2	3	4	5	6	7
Tid	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,∞]
Frekvens	9	16	51	72	79	11	12
Normal-sannsynlighet	0.0062	0.0606	0.2417	0.3829	0.2417	0.0606	0.0062

a) For videre studier av tiden det tar for barn å knyte skolissene, er forskerne interessert i å teste om de kan anta at den totale tiden det tar å knyte skolissene er normalfordelt med forventning 35 og standardavvik 10.

Beregn sannsynligheten, under den gitte antagelsen, for at den totale tiden det tar å knyte skolissene for et tilfeldig barn, ligger i intervallet $(30, 40]$. Sammenlign dette resultatet med tallet funnet i raden merket "Normal-sannsynlighet" og kolonnen merket "4".

Skriv ned nullhypotesen og den alternative hypotesen og utfør en hypotesetest for å test om dataene kommer fra en normalfordeling med forventning 35 og standardavvik 10. Baser denne testen på tabellen ovenfor. Bruk et 5% signifikansnivå.

Hva er konklusjonen fra denne testen?

Hvordan kunne du isteden teste nullhypotesen at den totale tiden det tar å knyte skolissene er normalfordelt, $N(\mu, \sigma)$, når μ og σ er uspesifisert? Du trenger bare kommentere med ord.

(Hint: Hvis X er binomisk fordelt, er $E(X) = n \cdot p$.)