NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4255 Anvendt statistikk**

Fagleg kontakt under eksamen: Anna Marie Holand

Tlf: 951 38 038

Eksamensdato: August 2016

Eksamenstid (frå–til):

Hjelpemiddelkode/Tillatne hjelpemiddel: C: Gult, stempla A4-ark med dine eigne handskrivne notatar, Tabeller og formler i statistikk (Tapir forlag/Fagbokforlaget). Bestemt kalkulator.

Annan informasjon:

- I utskrifta frå MINITAB er komma brukt som desimalskilleteikn.
- Signifikansnivå 5% skal brukast om ikke anna er spesifisert.
- Alle svar må grunngjevast.

Målform/språk: nynorsk

Sidetal: 10

Sidetal vedlegg: 0

Kontrollert av:

Dato Sign

Oppgåve 1 Behandling for reduksjon av stress

Det blei utført eit studie der tre ulike behandlingar mot reduksjon av stress vart sammenlikna. I studien vart 30 deltakarar tilfeldig delt inn i tre forskjellige behandlingsgrupper, gruppe A blei gjeve ei mental behandling, gruppe B blei gjeve behandling med fysisk trening, og gruppe C blei gjeve ei medisinsk behandling for å redusere stress. Ein poengskår mellom 1 og 5 vart gjeve etter at behandlinga var ferdig. Poengskåren angjev kor effektiv behandlinga var for å redusere deltakarnes stressnivå, der eit høgt poengskår indikerar høg effektivitet i å redusere stress.

Data og deskriptive mål for behandlingspoengskår for dei 3 forskjellige gruppene er gjeve 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 sammenlikne gruppen som fikk mental behandling (gruppe A) og gruppen som fikk behandling med fysisk trening (gruppe B).

a) La S_A og S_B vere standardavvikane for gruppe A og gruppe B og la desse vere estimatorar 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 berekne eit 95% konfidensintervall for $\frac{\sigma_A}{\sigma_B}$.

Basert på desse dataane, kan vi konkludere med at poengskårane for reduksjon i stress er forskjellige for dei to behandlingsgruppene (gruppe A og gruppe B)? Skriv ned nullhypotesen og den alternative hypotesen og utfør ein t-test for to grupper basert på dei deskriptive måla i tabell 2.

Bruk signifikansnivå $\alpha = 0.05$. Spesifiser kva for antakingar du gjer. Kva er konlusjonen frå testen?

b) Ein Wilcoxon rank-sum (Mann–Whitney) test blei utført på dataane frå gruppe A og gruppe B for å teste om poengskårane for reduksjon i stress er forskjellige for dei to behandlingsgruppene (gruppe A og gruppe B).

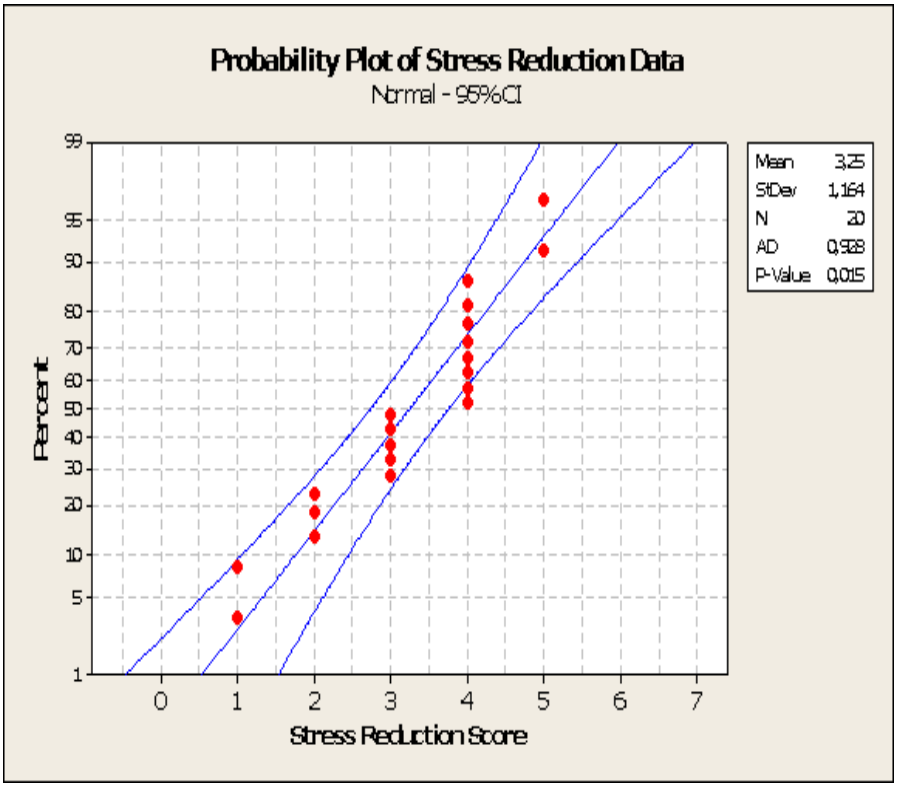
Skriv ned nullhypotesen og den alternative hypotesen for denne testen. Frå MINITAB-utskrifta i figur 1, er testobservatoren berekna til $W_1 = 116$. Vis korleis denne verdien kan finnast frå dataane i tabell 1. Forklar kort korleis verdien av W_1 vert brukt for å teste hypotesen.

Bruk signifikansnivå $\alpha = 0.05$.

Kva for antakelsar må du gjere i denne testen?
Kva for ein av testane i a) (andre halvdel) og b) syns du er mest passande for dataane? Du kan basere svaret på figur 2.

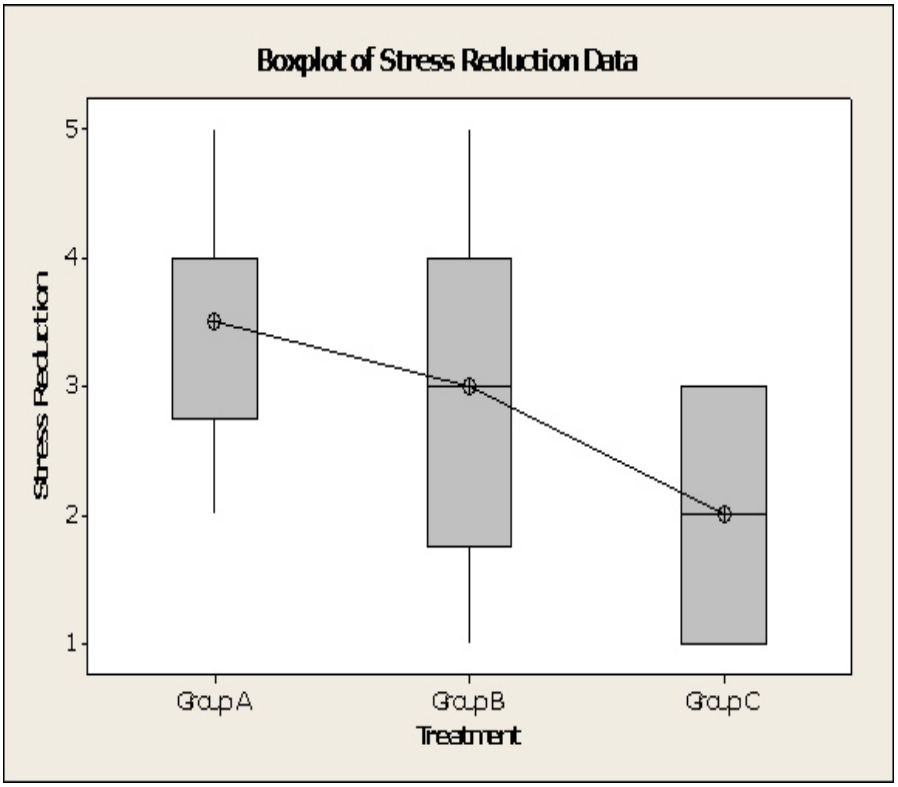
```
Mann-Whitney Test and CI

         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 frå ein Wilcoxon rank-sum test (Mann-Whitney) for datasettet for behandling for reduksjon av stress.



Figur 2: Eit normalplott basert på datasettet for behandling for reduksjon av stress, der dataane frå gruppe A og gruppe B er slått sammen.

I resten av oppgåva skal du bruke alle dei tre stressbehandlingsgruppene, A, B og C. Eit boksplott for dei tre behandlingsgruppene er gjeve 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 ein skilnad i effekten av stressreduksjon mellom dei tre behandlingsgruppene.

Berekn først kvadratsummane, SSA og SSE, ved å bruke dei deskriptive måla i tabell 2.

Utfør ei einvegs variansanalyse for å teste nullhypotesen

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

Bruk signifikansnivå $\alpha = 0.05$. Kva er antakingane du må gjere i denne testen? Kva er konklusjonen av testen?

d) Frå resultata i c) kom ein fram til at ein vil utforske vidare om nokre av skilnadane $\mu_A - \mu_B$, $\mu_A - \mu_C$ og $\mu_B - \mu_C$ er signifikant forskjellige frå 0.

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

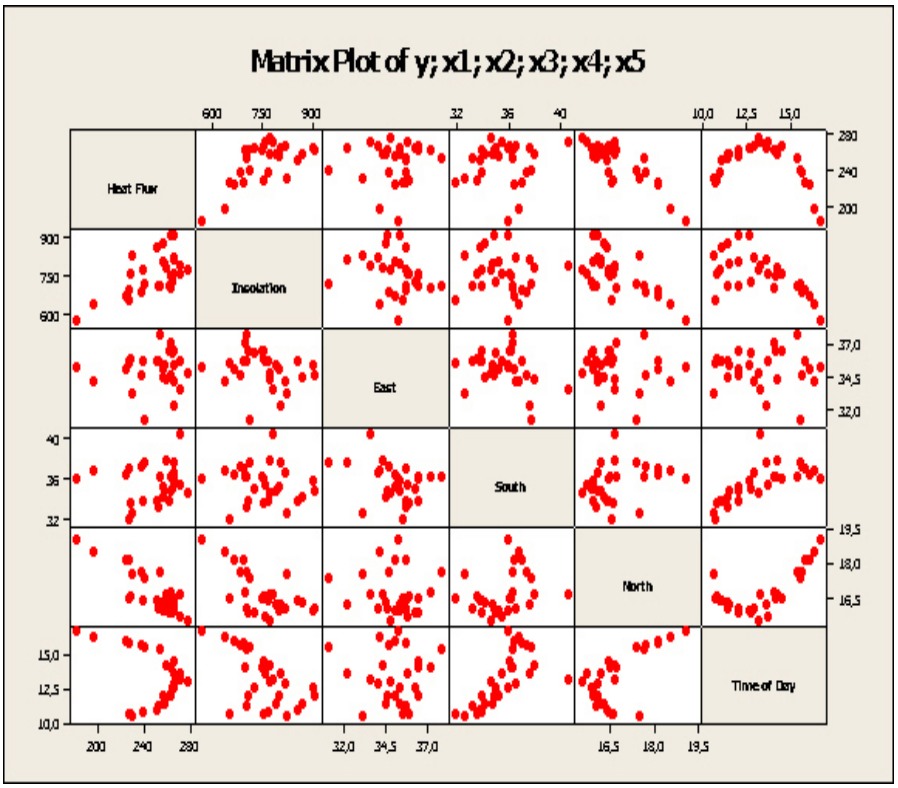
Oppgåve 2 Varmegjennomstrøyming

Ingeniørar som jobbar med energi er interessert i varmeenergi frå sola som del av utviklinga av solcelleenergi. Varmegjennomstrøyming (heat flux) er raten varmeenergi som vert overført gjennom ei gjeve overflate per eining tid. Varmegjennomstrøyming vert målt som ein del av ein energitest. Ein ingeniør er interessert i å bestemme korleis den totale varmegjennomstrøyminga vert predikert av variablane: innstråling, posisjonane i øst, sør og nord for brennpunktane, og tida på dagen. $n = 29$ varmeenergitestar vert tatt.

Følgjande beskriving er gjeve.

- y: den totale varmegjennomstrøyminga, i kilowatt
- x_1: innstrålingsverdien, i watt/m2
- x_2: posisjonen for brennpunkta i retning øst, i tommar
- x_3: posisjonen for brennpunkta i retning sør, i tommar
- x_4: posisjonen for brennpunkta i retning nord, i tommar
- x_5: tid på dagen dataane ble teken

Eit parvis spreiingsplott og ein korrelasjonsmatrise av variablane finst 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 spreiingsplott (øvre del) og parvis Pearson korrelasjon (nedre del) mellom variablane y, x_1, x_2, x_3, x_4 og x_5 i varmegjennomstrøymingsdatasettet.

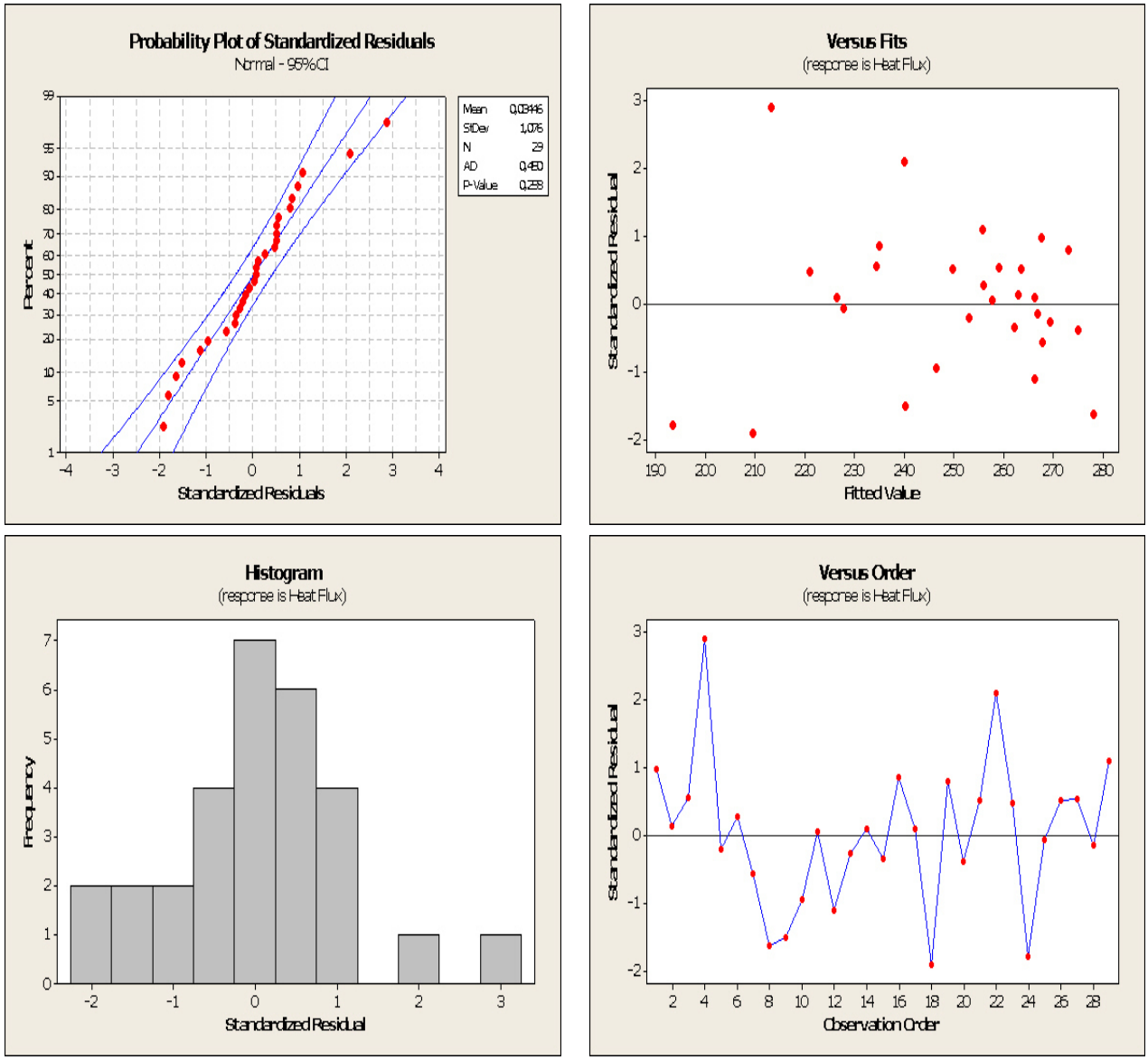
Ein multippel lineær regresjonsmodell vert tilpassa til dataane med y som respons og x_1, x_2, x_3, x_4 og x_5 som forklaringsvariabler. La (y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i} og x_{5i}) angje ein observasjon frå varmeenergitest i, der $i = 1, ..., 29$. Definer den fulle modellen (modell A):

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

der ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, ..., n$. MINITAB-utskrifta frå ei statistisk analyse av modell A finst i figur 5. Plott av standardiserte residual finst 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 frå tilpassinga av den lineære regresjonsmodellen A i varmegjennomstrøymingsdatasettet.



Figur 6: Residualplott (normal plott basert på standardiserte residual i øvre venstre panel, standardiserte residual mot tilpassa verdiar i øvre høgre panel, histogram basert på standardiserte residual i nedre venstre panel og standardiserte residual mot rekkjefølgja på observasjonane i nedre høgre panel) for regresjonsmodell A i varmegjennomstrøymingsdatasettet.

a) Skriv ned den estimerte regresjonslikninga.

Nå skal vi sjå 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?

Berekn R^2 og forklar korleis du kan tolke denne verdien.

Kva er eit passande estimat for σ? Gje den numeriske verdien av dette estimatet.

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

b) Kva for modellantakingar vert gjort i model A in likning (1)? Forklar og definer kva eit residual er.

Figur 6 visar nokon residualplott for model A. Indikerer residualplotta at desse antakingane for lineær regresjon er oppfylt? Grunngje svaret ved å kommentere kort på plottane i figur 6. Korleis ville dei ideelt sjå ut dersom modellen er korrekt?

Vi er no interessert i å samanlikne ulike regresjonsmodellar, der vi betraktar ulike kombinasjonar av forklaringsvariablane x_1, x_2, x_3, x_4 og x_5, i ein "best subset" regresjon. Anta at eit konstantledd, β_0, er med i regresjonsmodellen.

MINITAB-utskrifta frå tilpasninga av ulike modellar for dataane er presentert i figur 7. Kvar rad i figur 7 svarar til ein modell. Talet på forklaringsvariablar inkludert i kvar modell (i tillegg til konstantleddet, β_0) finn du i kolonnen med navn $Vars$. Dei to beste modellane for kvart tal på forklaringsvariablar er rapportert. X'a indikerar kva variablar som er funne 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 frå statistisk analyse av varmegjennomstrømningsdatasettet.

c) Basert på resultata i figur 7, kven av desse regressjonsmodellane ville du valgt som den "beste" modellen for datasettet? Grunngje valet ditt.

For å svare på dette spørsmålet skal du også forklare korleis R^2, R^2_{adj}, R^2_{pred}, Mallows C_p og S er definert og korleis du kan bruke dem til å sammenlikne dei ulike regresjonsmodellane.

Oppgåve 3 Tida det tar å knyte skolissar

Eit pilotstudie vert gjennomført der det blei målt tida det tek, i sekund, for barn på 9 år å knyte skolissane.

Totalt var talet barn på $n = 250$. Tida, gjeve i intervallar, og antall barn i kvart intervall finst i følgjande 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-sannsynet	0.0062	0.0606	0.2417	0.3829	0.2417	0.0606	0.0062

a) For vidare studiar av tida det tek for barn å knyte skolissene, er forskarne interessert i å teste om dei kan anta at den totale tida det tek å knyte skolissane er normalfordelt med forventning 35 og standardavvik 10.

Berekn sannsynet, under den gjeve antakinga, for at den totale tida det tek å knyte skolissane for eit tilfeldig barn, ligg i intervallet $(30, 40]$. Sammenlikn dette resultatet med talet funne i rada merka "Normal-sannsynet" og kolonna merka "4".

Skriv ned nullhypotesen og den alternative hypotesen og utfør ein hypotesetest for å test om dataane kjem frå ei normalfordeling med forventning 35 og standardavvik 10. Baser denne testen på tabellen ovanfor. Bruk eit 5% signifikansnivå.

Kva er konklusjonen frå denne testen?

Korleis kunne du i staden ha testa nullhypotesen at den totale tida det tek å knyte skolissane er normalfordelt, $N(\mu, \sigma)$, når μ og σ er uspesifisert? Du treng berre kommentere med ord.

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