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: 3. juni 2016

Eksamenstid (fra–til): 09:00-13:00

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: 9

Antall sider vedlegg: 0

Kontrollert av:

Dato Sign

Oppgave 1 Fisk og parasitter

I et eksperiment ble 141 fisker plassert i en stor tank. Fiskene ble klassifisert etter deres nivå av parasittinfeksjon, enten som ikke-infisert, lett infisert, eller høyt infisert. Noen av fiskene ble spist av rovfugler. Det er til parasittens fordel å være i en fisk som blir spist av en fugl ettersom dette gir muligheten til å infisere fuglen i parasittens neste livsstadie. Følgende kryss-tabell ble observert.

	Ikke-infisert	Lett infisert	Høyt infisert	Totalt
Spist	1	10	37	48
Ikke spist	49	35	9	93
Totalt	50	45	46	141

a) Forskerene som utførte forsøket ville undersøke om det å *bli spist eller ikke* og *nivå av parasittisk infeksjon* kan sees på som to avhengige hendelser?

Skriv ned nullhypothesen og den alternative hypotesen og utfør en hypotesetest basert på tabellen ovenfor. Bruk et 5% signifikansnivå.

Hva er konklusjonen fra testen?

Oppgave 2 Sigaretter

Den føderale handelskommisjonen i USA rangerer årlig varianter av sigaretter etter tjære-, nikotin-, og karbonmonoksidinnhold. Hvert av disse stoffene er farlige for helsen til en røyker. Tidligere studier har vist at økning i tjære og nikotininnhold i sigaretter er etterfulgt av en økning i karbonmonoksidmengden som blir sluppet ut fra sigarettrøyken.

I en studie ble følgende variabler målt for $n = 25$ sigarettmerker,

- y: Karbonmonoksidinnholdet (CO) (mg),
- x_1: Tjæreinnholdet (mg),
- x_2: Nikotininnholdet (mg), og
- x_3: Vekt (g).

Først ble tre separate enkle regresjonsmodeller tilpasset for å studere forholdet mellom innholdet av CO og hver av variablene x_1, x_2 and x_3:

$$y_i = \beta_{01} + \beta_1 x_{1i} + \epsilon_i \quad (1)$$
$$y_i = \beta_{02} + \beta_2 x_{2i} + \epsilon_i \quad (2)$$
$$y_i = \beta_{03} + \beta_3 x_{3i} + \epsilon_i \quad (3)$$

hvor ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, ..., n$.

MINITAB-utskrift fra en statistisk analyse finnes i Figur 1.

```
Simple regression for x1:
Predictor      Coef  SE Coef      T      P
Constant    1,4129   0,6482   2,18  0,040
x1         0,92813  0,05283  17,57  0,000

S = 1,11865   R-Sq = 93,3%   R-Sq(adj) = 93,0%
```

```
Simple regression for x2:
Predictor     Coef  SE Coef      T      P
Constant   -0,238    1,083  -0,22  0,828
x2         14,860    1,247  11,92  0,000

S = 1,58842   R-Sq = 86,6%   R-Sq(adj) = 86,0%
```

```
Simple regression for x3:
Predictor    Coef  SE Coef      T      P
Constant   -3,86    10,44  -0,37  0,715
x3         16,56    10,82   1,53  0,140

S = 4,12276   R-Sq = 9,6%   R-Sq(adj) = 5,5%
```

Figur 1: Utskrift fra tilpassing av de enkle regresjonsmodellene i ligning (1)-(3) for sigarettdatasettet.

a) Kommenter resultatene fra de enkle regresjonsmodellene i figur 1.

Vi vil nå fokusere på den enkle lineære regresjonsmodellen for x_2 i ligning (2) som er tilpasset i det midterste panelet i figur 1.

I den enkle lineære regresjonsmodellen for x_2 er en p-verdi gitt i raden merket x2. Forklar med ord hva denne p-verdien betyr.

Finn et 90% konfidensintervall for β_2 i den enkle lineære regresjonsmodellen for x_2.

Hva er et passende estimat for σ i den enkle lineære regresjonsmodellen for x_2?

Videre ble det utført en multippel regresjon med både x_1 og x_2 som kovariater.

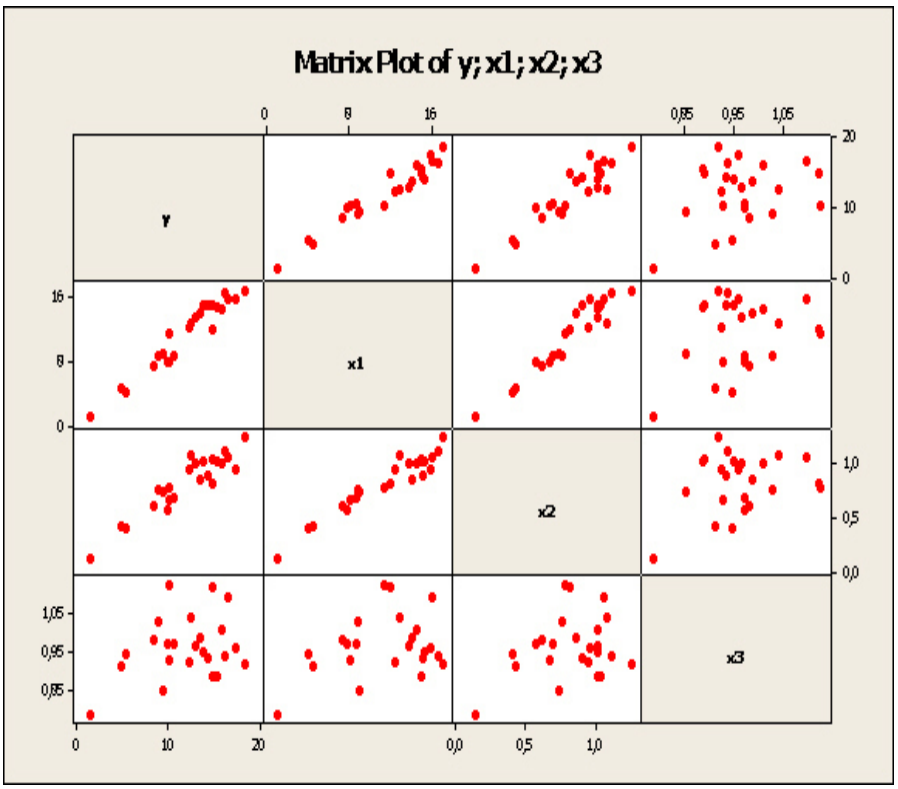
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{4}$$

hvor ϵ_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, ..., n$.

MINITAB-utskrift fra den tilpassede multiple regresjonsmodellen finnes i figur 2. Parvise spredningsplott for x_1, x_2, x_3 og y finnes i den øvre delen av figur 3 og parvise Pearson korrelasjoner finnes i nedre del av figur 3.

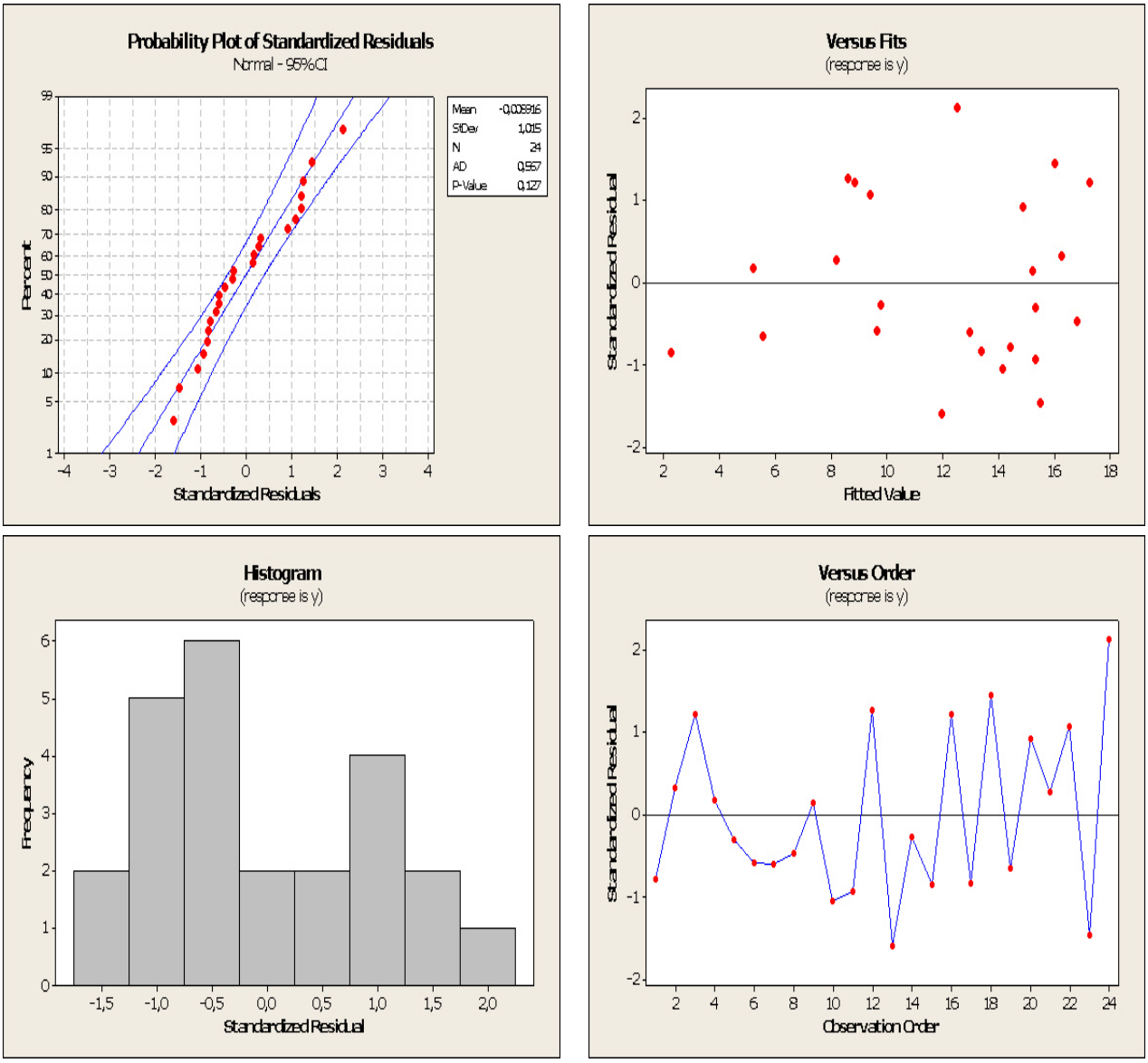
```
Predictor     Coef  SE Coef     T      P
Constant   1,3089   0,8483  1,54  0,138
x1         0,8918   0,1927  4,63  0,000
x2          0,629    3,203  0,20  0,846

S = 1,14392   R-Sq = 93,4%   R-Sq(adj) = 92,7%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        2  386,27  193,13  147,59  0,000
Residual Error   21   27,48    1,31
Total            23  413,75
```

Figur 2: Utskrift fra statistisk analyse av sigarettdataene for modellen i ligning (4).



	y	x1	x2
x1	0,966		
x2	0,931	0,960	
x3	0,310	0,284	0,286

Figur 3: Parvise spredningsplott (øvre del) og parvise Pearson korrelasjon (nedre del) mellom variablene y, x_1, x_2 og x_3 i sigarettdatasettet.



Figur 4: Residualplott (normalplott 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 regresjonsmodellen i ligning (4) for sigarettdatasettet.

b) Basert på modellen i ligning (4) hva er det predikerte CO innholdet når $x_1 = 10$ og $x_2 = 0.8$?

Basert på plottene i figur 4 og statistiske resultat av modelltilpassingen i figur 2, vil du si at modellen i ligning (4) er en god modell for dataene? Du må spesifisere hvilke egenskaper til tilpassingen av regresjonsmodellen og plottene du bruker for å komme frem til svaret ditt.

c) Ved tilpassing av en enkel lineær regresjon med bare nikotin (x_2) som kovariat, miderste panel i figur 1, fant vi at effekten av nikotin var signifikant på et 5% signifikansnivå, men i den multiple lineære regresjonen med tjære (x_1) og nikotin (x_2), figur 2, er nikotin ikke signifikant. Hva kan være årsaken til dette? Begrunn svaret.

Forklar begrepet *multikollinearitet*.

I MINITAB utskriften fra den tilpassete multiple regresjonsmodellen i figur 2 er det utført tre t-tester og en F-test. Forklar forskjellen mellom disse t-testene og F-testen.

Oppgave 3 Levetid til batteri

En produsent av batterier ville undersøke om levetiden til batteria er avhengig av de to faktorene *materialtype* og *driftstemperatur*. Tre materialtyper, kalt type 1, 2 og 3, og tre driftstemperaturer, Lav (-10°C), Medium (20°C) and Høy (45°C), ble undersøkt, og respons ble målt som den effektive levetiden av et batteri (timer, målt på en kontinuerlig skala). Forsøket ble utført ved å velge tilfeldig 12 batterier for hver materialtype, og så tilfeldig fordele batteriene til hver av de tre temperaturnivåene. Totalt 36 målinger ble tatt.

En to-veis variansanalysemodell (ANOVA) med samspill ble tilpasset til dataene og resultatet er gitt i tabell 1.

Source	DF	SS	MS	F value	p-value
Material	2	10684	?	7.9114	0.001976
Temperatur	2	39119	19559.4	28.9677	$1.909 \cdot 10^{-7}$
Material·Temperatur	4	?	2403.4	3.5595	?
Error	27	18231	675.2		
Total	?	?			

Tabell 1: Result fra to-veis ANOVA med samspill på batteridataene.

a) Hvilke antagelser ligger bak denne analysen?

Fem av verdiene i tabell 1 er erstattet med et spørsmålstegn (?). Regn ut tallverdier for hver av disse og forklar hva hvert av tallene betyr.

Er det en significant effekt av samspillsleddet Material·Temperature? Utfør en hypotesetest for å svare på dette spørsmålet. Skriv ned null hypotesen og den alterative hypothesen. Bruk et $\alpha = 0.05$ signifikansnivå.

Forklar viktige egenskaper ved resultatene fra denne to-veis ANOVAen og hvordan du vil gå videre med å analysere disse dataene?

Produsenten var interessert i å sammenligne levetiden til batteria for driftstemperaturene Medium and Høy for materialtype 3.

Det var $n_{Medium} = 4$ observasjoner for driftstemperatur Medium for materialtype 3. Gjennomsnittlig observert levetid for batteriene var $\bar{x}_{Medium} = 145.75$ og det empiriske standardavviket var $s_{Medium} = 22.54$. Videre var det $n_{\text{Høy}} = 4$ observasjoner for driftstemperatur Høy for materialtype 3. Gjennomsnittlig observert levetid for batteriene var $\bar{x}_{\text{Høy}} = 85.50$ og det empiriske standardavviket var $s_{\text{Høy}} = 19.28$.

b) Utfør en hypotesetest for å undersøke om forventet levetid for batterier ved de to driftstemperaturene Medium og Høy er ulike for materialtype 3. Skriv ned antagelsene du trenger å gjøre for å utføre denne testen.

La μ_{Medium} være forventet levetid for batterier ved driftstemperatur Medium for materialtype 3. Produsentene var interessert i forventet levetid for batterier på den naturlige logaritmiske skalaen, det vil si

$$\gamma = \ln(\mu_{Medium}).$$

c) Basert på det uavhengige tilfeldige utvalget av størrelse $n_{Medium} = 4$ fra driftstemperatur Medium for materialtype 3 foreslå en estimator, $\hat{\gamma}$, for γ.

Bruk tilnærmede metoder for å finne forventningsverdi og varians for denne estimatoren, det vil si, $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$. Bruk sammendraget av dataene gitt i teksten til å regne ut $\hat{\gamma}$ numerisk og gi estimert numerisk verdi for $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$.

Hint: Du kan benytte at $\frac{d}{dx}(\ln x) = \frac{1}{x}$.

Oppgave 4 Vaksineeffektivitet

En vaksine mot tyfoidfeber ble testet med den hensikt å kontrollere effektiviteten på vaksinen. Testen for effektivitet er målt som den biologiske aktiviteten til vaksinen. Hver uke ble tre prøver av vaksinen testet for dens effektivitet, over en periode på 13 uker.

La X_{ij} være målet på effektiviteten til vaksinen for prøve j, i uke i, hvor $j = 1, 2, 3$ og $i = 1, 2, ..., 13$. Videre, $\bar{X}_i = \frac{1}{3}\sum_{j=1}^{3} X_{ij}$, $S_i = \sqrt{\frac{1}{2}\sum_{j=1}^{3}(X_{ij} - \bar{X}_i)^2}$, $\bar{\bar{X}} = \frac{1}{13}\sum_{i=1}^{13} \bar{X}_i$, og $\bar{S} = \frac{1}{12}\sum_{i=1}^{13} S_i$.

Basert på disse 13 utvalgene, som vi antar er i kontroll, finner vi $\bar{\bar{x}} = 1.012$ og $\bar{s} = 0.168$.

a) Konstruer et S-chart og et $\bar{X}$-S-chart (med 3σ grenser).

Et nytt utvalg ble tatt, med $\bar{x} = 0.93$ og $s = 0.65$. Ser prosessen ut til å være i kontroll for dette utvalget? Grunngi svaret.

Table A.22 Factors for Constructing Control Charts

Obs. in Sample	Chart for Averages: Factors for Control Limits		Chart for Standard Deviations: Factors for Centerline		Chart for Standard Deviations: Factors for Control Limits				Chart for Ranges: Factors for Centerline		Chart for Ranges: Factors for Control Limits		
n	A_2	A_3	c_4	$1/c_4$	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_3	D_4
2	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.267
3	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	2.574
4	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	2.282
5	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	2.114
6	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	2.004
7	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.076	1.924
8	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.136	1.864
9	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.184	1.816
10	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.223	1.777
11	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.256	1.744
12	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.283	1.717
13	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	0.307	1.693
14	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	0.328	1.672
15	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	0.347	1.653
16	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	0.363	1.637
17	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	0.378	1.622
18	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	0.391	1.608
19	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	0.403	1.597
20	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	0.415	1.585
21	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	0.425	1.575
22	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	0.434	1.566
23	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	0.443	1.557
24	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	0.451	1.548
25	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	0.459	4.541

Figur 5: Table A22. Factors for constructing control charts.