Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen:
Mette Langaas (988 47 649)

BOKMÅL

EKSAMEN I TMA4255 ANVENDT STATISTIKK

Onsdag 8. august 2012
Tid: 9:00–13:00

Antall studiepoeng: 7.5.

Tillatte hjelpemidler: Alle trykte og håndskrevne hjelpemidler. Spesiell kalkulator.

Sensurfrist: 30. august 2012.

Eksamensresultatene annonseres fra `http://studweb.ntnu.no/`.

Merk deg følgende:

- Signifikansnivå 5% skal brukes hvis ikke annet er spesifisert.
- Alle svar må begrunnes.

Oppgave 1 Cannabisbruk blant ungdom i Norge

Statens institutt for rusmiddelforskning er ansvarlig for drift av statistikkdatabasen RusStat som har tilgjengelig statistikk om alkohol, narkotika og tobakk.

Vi skal nå se på data for bruk av cannabis blant ungdom for årene 2004–2008. I raden som heter "Totalt" er antall ungdommer som er spurt om cannabisbruk hvert år angitt.

	2004	2005	2006	2007	2008	Alle år
Brukt	222	261	206	357	281	1327
Ikke brukt	1432	1482	1365	2891	2394	9564
Totalt	1654	1743	1571	3248	2675	10891

a) Vil vil undersøke om det er en endring i frekvensen av cannabisbruk over årene 2004–2008. Skriv ned nullhypotesen og den alternative hypotesen og utfør en hypotesetest på grunnlag av informasjonen i tabellen over. For å forenkle beregningsbyrden kan det oppgis at χ^2-testobservatoren blir 27.7, og du trenger bare å vise hvordan du regner ut ett av de 10 leddene i summen.
Hva er din konklusjon basert på denne testen?

Oppgave 2 Prosessutvikling

Vi vil se på et designet eksperiment for å utvikle en etseprosess. Det er tre designfaktorer (A, B and C). De tre designfaktorene ble kjørt med to nivå hver.

- A: Avstanden mellom elektrodene. Lav: 0.80 cm. Høy: 1.20 cm.
- B: Gassflyten. Lav: 125. Høy: 200.
- C: Energi tilført katoden. Lav: 275 W. Høy: 325 W.

Reponsvariabelen er etseraten til prosessen (Å/m). Vi ønsker en høy verdi for etseraten. Det ble utført et 2^3 faktorielt design, og resultatene fra forsøket er presentert i tabell 1.

Run	A	B	C	Response
1	−1	−1	−1	550
2	1	−1	−1	669
3	−1	1	−1	633
4	1	1	−1	642
5	−1	−1	1	1037
6	1	−1	1	749
7	−1	1	1	1075
8	1	1	1	729

Tabell 1: Etse-eksperimentet med en replikat.

Fra disse resultatene har vi følgende estimater for eksperimentelt design (DOE) faktor effekter.

A	B	C	AB	AC	BC	ABC
−126.5	*	274.0	−42.0	−190.5	−9.5	13.0

a) Fyll inn det manglende effektestimatet for faktor B. Lag et hovedeffekt-plott (main effects plot) for faktor B, og forklar med ord hvordan du kan tolke hovedeffekten av faktor B.

Dette eksperimentet ble gjentatt, slik at hver faktorkombinasjon ble kjørt totalt to ganger, dvs. $n = 16$. Resultatene fra en multippel lineær regresjon der konstantledd, A, B, C, AB, AC, BC og ABC ble brukt som kovariater, finner du i tabellen under. Vi kaller denne den *fulle* regresjonsmodellen. Merk at estimatene ("Estimate") er de estimerte regresjonskoeffisientene og ikke effektestimatene.

	Estimate	Std. Error	t value	p-value
Intercept	776.062	11.865	65.406	3.32e-12
A	−50.812	11.865	−4.282	0.002679
B	3.688	11.865	*	0.763911
C	153.062	11.865	12.900	1.23e-06
AB	−12.437	11.865	−1.048	0.325168
AC	−76.812	11.865	−6.474	0.000193
BC	−1.062	11.865	−0.090	0.930849
ABC	2.813	11.865	0.237	0.818586
Residual standard error: 47.46 on 8 degrees of freedom				

b) Hva er angitt i kolonnnen "Std. Error", og hvorfor er "Std. Error" den samme for alle kovariatene?

Nå skal vi se på andre rad i tabellen, som omhandler faktor B. Hva er sammenhengen mellom en estimert koeffisient for B og en estimert effekt for B? Regn ut det manglende tallet for t-observatoren i tabellen. Hvilke hypoteser bygger p-verdien på, og hva blir konklusjonen for faktor B?

Hvilke kovariater er signifikante i modellen?

c) Anta nå at vi bruker en regresjonsmodell med konstantledd og kovariatene A, C og AC. Resultatene fra tilpassing av denne modellen finner du i tabellen under. Vi kaller dette den *reduserte* regresjonsmodellen.

	Estimate	Std. Error	t value	p-value
Intercept	776.06	10.42	74.458	<2e-16
A	−50.81	10.42	−4.875	0.000382
C	153.06	10.42	14.685	4.95e-09
AC	−76.81	10.42	−7.370	8.62e-06

Residual standard error: 41.69 on 12 degrees of freedom

Hvorfor er de estimerte koeffisienten for A, C og AC de samme for den fulle og den reduserte modellen? Hvorfor har standardavvikene endret seg fra full til redusert modell?

Hva vil du foreslå som optimale valg av nivå (lavt eller høyt) for hver av disse to faktorene når målet er å bruke den tilpassede regresjonsmodellen til å finne kombinasjonen med høyest estimert etserate?

Regn ut et 95% prediksjonsintervall for etseraten for disse valgte nivåene for A og C.

d) Anta nå at i en pilotstudie med tre faktorer så ble kjøringene 1, 4, 6 og 7 fra tabell 1 utført.

Hvilken type eksperiment er dette?
Hva er generatoren og den definerende relasjonen for eksperimentet?
Skriv ned aliasstrukturen for eksperimentet.
Hvilken resolusjon har eksperimentet?

Oppgave 3 Genaktivitet

Vi skal se på resultater fra en biologisk studie der målet var å sammenligne genaktivitet for tre gener (B_1, B_2, og B_3) i tre ulike cellepopulasjoner (A_1, A_2 og A_3). Genaktiviteten ble målt ved polymerase kjedereaksjonsmetoden. Målingene av genaktivitet er kontinuerlige og analyseres på logaritmisk skala (base 2). Totalt ble det gjort $n = 45$ målinger.

a) En toveis variansanalysemodell (ANOVA) med samspill be tilpasset til dataene, og gav følgende resultater.

Source	Df	Sum Sq	Mean Sq	F value	p-value
A	2	3.240	*	5.935	*
B	2	253.161	126.581	463.768	<2e-16
AB	4	*	30.645	112.277	<2e-16
Error	36	9.826	0.273		
Total	*	388.807	8.837		

Fire av tallene i tabellen er blitt erstattet med en asterisk (stjerne). Forklar hva kolonneoverskriftene betyr og regn ut numeriske verdier for de fire manglende tallene. Er samspillsleddet AB signifikant? Hvordan vil du gå videre med å analysere disse dataene?

I resten av oppgaven vil vi bare se på resultater for cellepopulasjon A_1.

b) Biologen som har utført studien vil sammenligne genaktiviteten for gen B_1 og B_2 for cellepopulasjon A_1. Det ble gjort $n_1 = 3$ observasjoner av gen B_1 i cellepopulasjon A_1 og gjennomsnittet var $\bar{x} = -29.1$ og det empiriske standardavviket $s_1 = 0.34$. Videre var det $n_2 = 9$ observasjoner for gen B_2 i cellepopulasjon A_1 og gjennomsnittet var $\bar{y} = -33.7$ og det empiriske standardavviket $s_2 = 0.42$.

Utfør en test for å undersøke om genaktiviteten for de to genene er ulik i denne cellepopulasjonen. Skriv ned antagelsene du trenger å gjøre for å utføre denne testen.

c) La μ_1 være forventet genaktivitet på logaritmisk skala (base 2) for gen B_1 i cellepopulasjon A_1, og la μ_2 være forventet genaktivitet på logaritmisk skala (base 2) for gen B_2 i cellepopulasjon A_1. Biologen er interessert i forventet *fold change* mellom aktiviteten til de to genene på originalskala, det vil si

$$\gamma = \frac{2^{\mu_1}}{2^{\mu_2}}.$$

Basert på to tilfeldige utvalg av størrelse n_1 og n_2 fra de to gruppene foreslå en estimator, $\hat{\gamma}$, for γ.

Bruk tilnærmede metoder for å finne forventingsverdi og varians for denne estimatoren, det vil si, $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$. Bruk utvalgsstørrelsene, gjennomsnittene og de empiriske standardavvikene gitt i b) til å regne ut numerisk verdi for $\hat{\gamma}$ og til å gi estimerte numeriske verdier for $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$.

Hint: Du kan benytte at $2^x = \exp(x \ln 2)$, der ln er den naturlige logaritmen.