UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2
Eksamensdag:	Mandag 4. juni 2007.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabeller for χ^2, t og F fordelingene.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/ STK1110 og for STK1120.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

For å sammenlikne tre ulike typer investeringsfond (som vi her vil kalle A, B og C), ble 2000 kroner investert i 18 ulike fond, 6 av hver type, i en periode på 5 år. For hver investering ble gevinsten (i kroner) registrert og er gitt i tabellen nedenfor.

A	B	C
13288	15738	14790
12782	14249	13827
12812	12369	13680
11713	12822	13150
11201	12117	12669
12233	12605	14267

Vi vil anta en modell

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, ..., I, j = 1, ..., J \tag{1}$$

for disse data. De vanlige antagelser vil gjelde for e_{ij}-ene.
En Anova tabell for analyse av disse data er gitt nedenfor.

(Fortsettes side 2.)

```
Analysis of Variance Table

Response: y
          Df   Sum Sq  Mean Sq F value  Pr(>F)
fund       2  6134625  3067312  2.9258 0.08455
Residuals 15 15725266  1048351
```

(a) Sett opp den hypotesen som testes.

Forklar hvordan du kan bruke tallene i Anova tabellen for å teste hypotesen og formulér en konklusjon på testen når du bruker signifikansnivå 0.05.

Vi vil nå diskutere mer generelt modellen (1). Anta vi ønsker å teste hypoteser på formen

$$H_0 : \sum_{i=1}^{I} c_i \alpha_i = 0 \tag{2}$$

der $\sum_{i=1}^{I} c_i = 0$. $C = \sum_{i=1}^{I} c_i \alpha_i$ kalles gjerne en *kontrast*.

(b) Definer $\widehat{C} = \sum_{i=1}^{I} c_i \overline{Y}_{i\cdot}$ der $\overline{Y}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$. Finn forventningen til $\widehat{C}$ og vis at denne forventningen blir lik 0 under H_0.

Vis også at

$$E[\widehat{C}^2] = C^2 + \text{var}[\widehat{C}].$$

(c) Finn variansen til $\widehat{C}$ og vis at den kan skrives som $\text{var}[\widehat{C}] = \sigma^2 K_C$ for en konstant K_C.

Definer $SS_C = \frac{\widehat{C}^2}{K_C}$. Argumenter for at

$$F_C = \frac{SS_C/1}{SS_W/(IJ-J)}, \quad \text{der} \quad SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i\cdot})^2.$$

er en fornuftig testobservator for å teste H_0. For hva slags verdier (store og/eller små) av F_C bør H_0 forkastes?

(d) Vis at SS_C/σ^2 blir χ^2-fordelt med 1 frihetsgrad og at SS_C er uavhengig av SS_W.

Argumenter også for at F_C er F-fordelt med 1 og $IJ - J$ frihetsgrader.

(e) Returner nå til investeringsdataene. Anta vi ønsker å teste om den forventede verdien for den første typen investeringsfond er lik gjennomsnittet av de forventede verdier for de to andre typer investeringsfond.

(Fortsettes side 3.)

Spesifisér dette som en test av en hypotese på formen (2).

Den tilhørende F_C verdi blir i dette tilfellet 5.362 (dette behøver du *ikke* å regne ut). Utfør testen og formulér en konklusjon når du igjen bruker signifikansnivå 0.05.

Diskutér konklusjonen i forhold til den konklusjon du fikk i (a).

Oppgave 2.

Genetisk variabilitet er antatt å være en sentral faktor i overlevelse av arter. Idéen er at populasjoner med mange "forskjellige" individer vil ha bedre sjanser for å klare seg i nye omgivelser. Tabellen nedenfor viser resultatene av et studie designet for å teste hypotesen eksperimentelt (hentet fra Larsen og Marx: *An Introduction to Mathematical Statistics and Its Applications*). To populasjoner av bananfluer, en som var kryssavlet (populasjon A med stor genetisk variabilitet) og en som var innavlet (populasjon B med mindre genetisk variabilitet) ble puttet i forseglede beholdere der mat og rom ble holdt til et minimum. Hver hundrede dag ble antall levende fluer i hver populasjon registert. Merk at dette vil være antall fluer i de neste generasjoner da bananfluer har levealder godt under 100 dager.

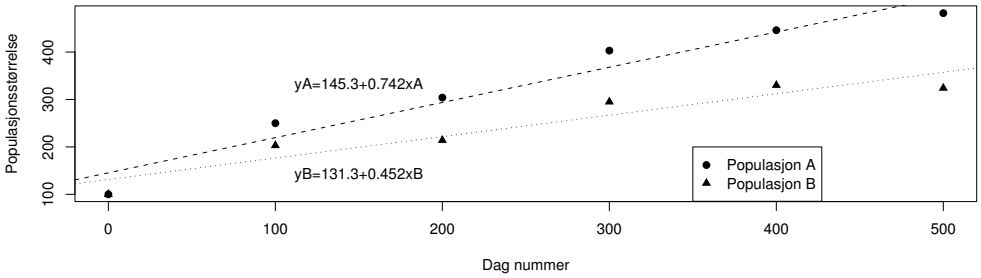
Dag nummer ($x^A = x^B$)	Populasjon A (y^A)	Populasjon B (y^B)
0	100	100
100	250	203
200	304	214
300	403	295
400	446	330
500	482	324

I figuren nedenfor er dataene plottet sammen med regresjonslinjer for modellene

$$\begin{aligned} y_i^A &= \beta_0^A + \beta_1^A x_i^A + e_i^A, \quad i = 1, ..., 6; \\ y_i^B &= \beta_0^B + \beta_1^B x_i^B + e_i^B, \quad i = 1, ..., 6. \end{aligned} \tag{3}$$

For begge populasjonene antas støyleddene å være uavhengige og normalfordelte med forventning 0 og felles varians σ^2. Det antas også uavhengighet mellom de to populasjonene.

(Fortsettes side 4.)



β_1-parametrene beskriver endringer i populasjonsstørrelser og kan brukes som indikatorer på overlevelsesmulighetene til populasjonene. Vi vil i denne oppgaven derfor se på hvordan vi kan teste

$H_0 : \beta_1^A = \beta_1^B$ mot $H_A : \beta_1^A > \beta_1^B$.

(a) La $\hat{\beta}_1^A$ og $\hat{\beta}_1^B$ være minste kvadraters estimatene til henholdsvis β_1^A og β_1^B under de to lineære regresjonsmodellene.

Hva er forventningen til $\hat{\beta}_1^A - \hat{\beta}_1^B$ under modell (3)?

Hva blir variansen til $\hat{\beta}_1^A - \hat{\beta}_1^B$ under den samme modellen?

(Du kan her bruke at $\text{var}[\hat{\beta}_1^A] = \text{var}[\hat{\beta}_1^B] = \frac{\sigma^2}{\sum_{i=1}^6 (x_i - \bar{x})^2}$ uten å utlede dette.)

Argumenter for at

$$T = \frac{\hat{\beta}_1^A - \hat{\beta}_1^B}{s_{\hat{\beta}_1^A - \hat{\beta}_1^B}} \tag{4}$$

er en fornuftig test-observator å bruke for å teste H_0 mot H_A under modell (3). (Her er $s_{\hat{\beta}_1^A - \hat{\beta}_1^B}$ standardfeilen til $\hat{\beta}_1^A - \hat{\beta}_1^B$.)

En kan vise at under modell (3) og H_0 så vil T følge en t-fordeling med 8 frihetsgrader (dette behøver du ikke å vise).

(b) For de gitte data blir $T = 2.49$. Bruk dette til å utføre en test på H_0 og formuler en konklusjon.

(c) Anta nå vi er noe usikre på antagelsene som ligger til grunn for modellene. Diskuter hvordan man kan bruke bootstrapping til å konstruere konfidensintervall for $\beta_1^A - \beta_1^B$.

Testprosedyren ovenfor er basert på separate regresjonsmodeller for hver av de to populasjonene. Et alternativ vil være å bruke en felles modell

$$y_i = \begin{cases} \beta_0 + \beta_1^A x_i + e_i & \text{for } i \text{ en observasjon fra populasjon A} \\ \beta_0 + \beta_1^B x_i + e_i & \text{for } i \text{ en observasjon fra populasjon B} \end{cases} \tag{5}$$

der i nå er en indeks som løper over alle 12 observasjoner.

(Fortsettes side 5.)

(d) Skriv modellen (5) som en multippel lineær regresjonsmodell med passende valg av forklaringsvariable.

Argumenter for hvorfor dette kan være en rimelig modell å bruke i den aktuelle situasjonen.

Tabellen nedenfor viser utskrift fra en regresjonsanalyse for fellesmodellen (5). I tillegg får du oppgitt at den estimerte korrelasjonen mellom $\hat{\beta}_1^A$ og $\hat{\beta}_1^B$ er 0.52.

	Estimate	Std. Error	t value
β_0	138.33333	16.76581	8.251
β_1^A	0.76109	0.06358	11.971
β_1^B	0.43291	0.06358	6.809

T-observatoren gitt i (4) kan brukes for å teste H_0 også i dette tilfellet, men nye verdier for estimatet av differansen og dets standardfeil må da brukes. En kan vise at T under modellen (5) og under H_0 blir t-fordelt med 9 frihetsgrader (dette behøver du ikke å vise). Den observerte verdien for T blir under den nye modellen lik 5.268.

(e) Gi en verbal begrunnelse for hvorfor vi nå ender opp med 9 frihetsgrader i stedet for 8 som vi hadde da vi brukte separate regresjonsmodeller.

Hva blir konklusjonen av å teste H_0 basert på modellen (5)?

Hvorfor er det rimelig at vi får en mindre P-verdi i dette tilfellet?

Oppgave 3.

Tabellen nedenfor viser hvor mange som overlevde forliset av *Titanic* fordelt på om de var mannskap ombord eller 1., 2. eller 3. klasses passasjerer (tallene i parentes viser E_{ij}-verdiene som brukes i en kji-kvadrattest på disse data). Du får også oppgitt at kji-kvadrat testobservatoren X^2 er 187.79.

	Mannskap	1. klasse	2. klasse	3. klasse	Totalt
Overlevd	212	202	118	178	710
	(285.48)	(104.84)	(91.94)	(227.74)	
Død	673	123	167	528	1491
	(599.52)	(220.16)	(193.06)	(478.26)	
Totalt	885	325	285	706	2201

Vi ønsker å undersøke om sjansen for å overleve er den samme uansett status på skipet.
Formuler en passende null hypotese for teste dette.
Utfør testen og formuler en konklusjon. Hva er de mest slående avvik fra null hypotesen?

SLUTT