UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Innføring i anvendt statistikk - FASIT
Eksamensdag:	Mandag 4. juni 2007.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 4 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.

(a) Hypotese: $H_0 : \alpha_i = 0, \forall i$.

Under H_0 er F F-fordelt med 2 og 15 frihetsgrader. P-verdien gir sannsynligheten for å få den observerte verdien eller mer ekstreme verdier og bør være liten for å forkaste H_0. Siden P-verdien her er større enn 0.05, forkaster vi ikke hypotesen.

Alternativt kan en velge et signifikansnivå, f.eks. $\alpha = 0.05$, og forkaste hvis F er større enn øvre α kvantil i tilhørende F-fordeling. Her er $F_{0.95}(2, 15) = 3.68$ og siden den observerte F ikke er større, forkaster vi ikke hypotesen.

Konklusjon: Data gir ikke grunnlag for å forkaste hypotesen om at de ulike fondstyper gir forkjellig gevinst.

(Fortsettes side 2.)

(b) Vi har $E\widehat{Y}_{i\cdot} = \mu + \alpha_i$ som gir at

$$\begin{aligned} E\widehat{C} =& \sum_{i=1}^{I} c_i E\widehat{Y}_{i\cdot} = \sum_{i=1}^{I} c_i[\mu + \alpha_i] \\ =& \mu \sum_{i=1}^{I} c_i + \sum_{i=1}^{I} c_i\alpha_i = \sum_{i=1}^{I} c_i\alpha_i \stackrel{H_0}{=} 0 \end{aligned}$$

2. del av oppgaven følger direkte av at $\widehat{C}$ er forventingsrett for C og av definisjonen for varians.

(c) Vi har

$$\begin{aligned} \text{var}[\widehat{C}] =& \text{var}[\sum_{i=1}^{I} c_i\overline{Y}_{i\cdot}] \stackrel{uavh}{=} \sum_{i=1}^{I} \text{var}[c_i\overline{Y}_{i\cdot}] \\ =& \sum_{i=1}^{I} c_i^2 \text{var}[\overline{Y}_{i\cdot}] = \sum_{i=1}^{I} c_i^2 \frac{\sigma^2}{J} = \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{J} \end{aligned}$$

dvs $K_C = \sum_{i=1}^{I} \frac{c_i^2}{J}$.

Under H_0 er $E[\widehat{C}^2] = \text{var}[\widehat{C}]$ og dermed er $E[SS_C] = \sigma^2$. Samtidig er $E[SS_W/(IJ - J)] = \sigma^2$ slik at under H_0 vil F ligge i nærheten av 1.

Utenfor H_0 vil telleren bli større (et tillegg på C^2). Det er derfor naturlig å forkaste H_0 når F_C er stor.

(d) $\widehat{C}$ er normalfordelt siden den er en lineær kombinasjon av normalfordelte variable. $\widehat{C}/\sigma\sqrt{K_C}$ er standard normalfordelt og dermed er $\widehat{C}^2/\sigma^2 K_C$ χ^2-fordelt med 1 frihetsgrad. Siden $\overline{Y}_{i\cdot}$ er uavhengig av $\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i\cdot})^2$, blir SS_C uavhengig av SS_W.

Dermed blir F_C forholdet mellom to uavhengige χ^2-fordelte variable og selv F-fordelt med frihetsgrad er 1 og $IJ - J$.

(e) Vi ønsker å teste om $H_0 : \mu + \alpha_1 = \frac{1}{2}(\mu + \alpha_2 + \mu + \alpha_3)$ som kan omformuleres til $H_0 : \alpha_1 - \frac{1}{2}\alpha_2 - \frac{1}{2}\alpha_3 = 0$. Dette svarer til $c_1 = 1$ og $c_2 = c_3 = -\frac{1}{2}$.

Frihetsgrader blir her 1 og 15. $F_{0.95}(1, 15) = 4.54$ mens $F_{0.975}(1, 15) = 6.20$. Siden den observerte F_C ligger mellom disse, må den tilhørende P-verdi ligge mellom 0.025 og 0.05. Det er dermed grunnlag for å forkaste hypotesen på 0.05 nivå og påstå at det *er* forskjell mellom den første typen investeringsfond og gjennomsnittet av de to andre.

Vi får her en litt annen konklusjon enn i (a). Dette skyldes at vi nå prøver å teste en enklere hypotese og dermed får bedre styrke. (Dette siste vil være noe vi ikke har diskutert særlig i kurset).

(Fortsettes side 3.)

Oppgave 2.

(a) Siden regresjonsparametrene begge er forventningsrette, blir $E[\hat{\beta}_1^A - \hat{\beta}_1^B] = \beta_1^A - \beta_1^B$.

Vi har at $\text{var}(\hat{\beta}_1^A) = \text{var}(\hat{\beta}_1^B) = \frac{\sigma^2}{\sum_{i=1}^6 (x_i - \overline{x})^2}$. Pga uavhengighet mellom de ulike populasjonene, blir

$$\text{var}(\hat{\beta}_1^A - \hat{\beta}_1^B) = \frac{2\sigma^2}{\sum_{i=1}^6 (x_i - \overline{x})^2}$$

(b) Forkaster når T er stor. Vi har at $t_{0.975}(8) = 2.306$ mens $t_{0.99}(8) = 2.896$ som tilsier at P-verdien for testen ligger mellom 0.025 og 0.05. Det er derfor grunnlag for å forkaste hypotesen, dog ikke veldig sterkt bevis.

(c) Vi kan utføre Bootstrapping ved å simulere data fra en estimert modell $\widehat{F}$ og for hver av de simulerte datasettene estimere β-parametrene. Dette gir oss også simulerte estimater på differansen $\beta_1^A - \beta_1^B$. Hvis vi lar D_b^* være den estimerte differansen fra bootstrap sample b, vil et standard bootstrap konfidensintervall for $\beta_1^A - \beta_1^B$ være

$$[2 * (\hat{\beta}_1^A - \hat{\beta}_1^B) - \overline{D}^*, 2 * (\hat{\beta}_1^A - \hat{\beta}_1^B) - \underline{D}^*]$$

der $\underline{D}^*$ og $\overline{D}^*$ er nedre og øvre kvantiler i den empiriske fordelingen til D_b^*-ene.

Her er det naturlig å oppfatte forklaringsvariablene som faste. Det er derfor naturlig å bruke

$$\begin{aligned} y_i^{A,*} &= \hat{\beta}_0^A + \hat{\beta}_1^A x_i^A + e_i^{A,*} \\ y_i^{B,*} &= \hat{\beta}_0^B + \hat{\beta}_1^B x_i^B + e_i^{B,*} \end{aligned}$$

der $e_i^{A,*}, e_i^{B,*}$ enten kan trekkes fra residualene fra modellene (ikke-parametrisk bootstrapping) eller fra normalfordeling (parametrisk bootstrap).

I begge tilfeller kan en velge om en vil la Bootstrap-fordelingen til støyleddene være felles for begge modellene eller ha separate fordelinger.

(d) Vi kan skrive modellen som

$$y_i = \beta_0 + \beta_1^A x_i z_i + \beta_1^B x_i (1 - z_i) + e_i$$

der $z_i = 1$ hvis observasjon i kommer fra populasjon A og 0 ellers.

I dette tilfellet starter vi med samme størrelse på populasjonene slik at det er rimlig å si at konstantleddet er felles (en kunne faktisk argumentere for at en bør sette $\beta_0 = 100$). Siden også støyleddene er antatt å ha samme varians, er det rimlig å estimere disse basert på alle data.

(Fortsettes side 4.)

(e) Siden vi med den felles modellen har én mindre parameter å estimere, får vi en ekstra frihetsgrad.

Vi har at $t_{0.995}(9) = 3.250$. Siden den observerte T er større enn denne, vil P-verdien være mindre enn 0.005. Dataene gir derfor et klart grunnlag for å forkaste H_0, dvs til å påstå at populasjonen med større genetisk variabilitet har en høyere vekstrate enn den andre populasjonen.

Et større antall frihetsgrader fører til større styrke i testen. Videre har vi med den felles modellen nå en positiv korrelasjon mellom $\hat{\beta}_1^A$ og $\hat{\beta}_1^B$ som gør at standardfeilen for differansen blir mindre. Dette gir oss en større T-verdi.

Oppgave 3.

Vi ønsker her å teste on sannsynligheten for å overleve er den samme for de ulike statusene passasjerene har. Det er derfor rimlig å bruke en test om homogenitet, dvs

$H_0 : \pi_{ij} = \pi_i, \quad$ for alle i, j.

Antall frihetsgrader blir her $(4-1)*(2-1) = 3$. $\chi^2_{0.995}(3) = 12.84$, som den observerte X^2 klart overstiger. Deter derfor en klar indikasjon på at status hadde betydning for overlevelsessannsynlighet ved *Titanic* katastrofen.
En har $O_{ij} - E_{ij}$ er

-73.484	97.161	26.065	-49.742
73.484	-97.161	-26.065	49.742

mens $(O_{ij} - E_{ij})^2/E_{ij}$ er

18.915	90.046	7.390	10.864
9.007	42.879	3.519	5.173

som viser at det først og fremst er den større overlevelsesraten i 1. klasse passasjerene som bidrar til den store X^2. Men det er også en god del høyere dødsrate innen manskapet enn forventet.

SLUTT