UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — FASIT.
Eksamensdag:	Tirsdag 11. desember 2012.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over Poisson-, normal-, t-, og F-fordeling.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110.

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

Oppgave 1.

a) Dette er en standard oppgave om sammenligning av to andeler, kapittel 10.4. Se deler av side 507-509 for utledning (large sample test procedure).

b) Dette er også standard, se side 512 i boken. Uttrykket for variansen blir forskjellig fordi vi i a) har antatt at H_0 er sann, dvs. $p_1 = p_2$.

c) La $\hat{p}_1 = 102/200 = 0.51$, $\hat{p}_2 = 248/400 = 0.62$ og $\hat{p} = (102 + 248)/(200 + 400) = 0.583$. Testobservator i a) får verdi $z = -2.576$. Med tosidig alternativ og $\alpha = 0.05$ må vi sjekke om $z > z_{0.025} = 1.96$ eller $z < -z_{0.025} = -1.96$. Siden observert z er mindre enn -1.96 forkaster vi H_0 og konkluderer på nivå 0.05 at det er signifikant forskjell i andel fulltidsstudenter.
Innsatt i formelen for konfidensintervallet får vi (-0.194, -0.026). At 0 ikke ligger i dette intervallet stemmer med at vi fikk forkastning av hypotesen om at $p_1 = p_2$.

Oppgave 2.

a) Vi setter score kvinner: $X_i \sim N(\mu_1, \sigma^2)$, $i = 1, ..., n$ og score menn: $Y_i \sim N(\mu_2, \sigma^2)$, $i = 1, ..., m$. Hypotesene blir

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_a : \mu_1 - \mu_2 < 0$$

Baserer oss på $\bar{X} \sim N(\mu_1, \sigma^2/n)$ og $\bar{Y} \sim N(\mu_2, \sigma^2/m)$. Fordi vi antar lik varians bruker vi $S_p = S_{pooled} = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m}$.

(Fortsettes på side 2.)

Testobservator

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2},$$

dvs. student-t-fordelt med $n + m - 2$ frihetsgrader. Med $n = 10$ og $n = 9$ blir dette 17 frihetsgrader. Fra R-utskriften finner vi observert verdi for t som t-statistic=- 0.959 og tilhørende p-verdi = 0.1755. Med nivå $\alpha = 0.05$ har vi ingen grunn til å forkaste H_0.

b) Hypotesene blir her $H_0 : \sigma_1^2 = \sigma_2^2$ mot $H_a : \sigma_1^2 \neq \sigma_2^2$. Vi vet at

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n-1}$$
$$\frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2_{m-1}$$

Testobservator

$$f = \frac{(n-1)S_1^2/\sigma_1^2(n-1)}{(m-1)S_2^2/\sigma_2^2(m-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1,m-1}$$

dvs. F-fordelt med $n - 1$ og $m - 1$ frihetsgrader. Innsatt n og m blir dette 9 og 8 frihetsgrader. Under H_0 er $f = \frac{S_1^2}{S_2^2} = 1.6688$ og vi finner p-verdi = 0.4822 fra utskriften. Med nivå $\alpha = 0.1$ har vi ingen grunn til å forkaste H_0.

c) Vi har modellen

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

Med indikatorvariabelen $x_2 = 1$ for kvinner og $x_2 = 0$ for menn, får vi modellene
Kvinner: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 + \beta_3 x_{1i} + \epsilon_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_{1i} + \epsilon_i$
Menn: $y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$
β_2 blir dermed forskjellen mellom kvinner og menn i grunnleggende lønnstillegg, og β_3 blir forskjellen mellom kvinner og menn i uttelling for score (dvs. forskjellen i stigningstall).

d) $\widehat{\beta}_2 = -31.123$, noe som betyr at kvinner typisk får ca. 31$ mindre i lønnstillegg i utgangspunktet. Men her er p-verdi = 0.5293, så denne forskjellen er ikke signifikant. For uttelling for score, derimot, har vi klare resultater: $\widehat{\beta}_3 = -3.96$, som betyr at stigningstallet, som sier hvor mye lønnsøkning man kan forvente for hvert ekstra poeng for arbeidsinnsats, er betraktelig lavere for kvinner enn for menn. Her er p-verdi= 0.00029 og forskjellen signifikant på alle rimelige nivåer.

e) Vi har at $\widehat{\beta}_3 \sim N(\beta_3, \sigma^2_{\widehat{\beta}_3})$ og videre

$$\frac{\widehat{\beta}_3 - \beta_3}{S_{\widehat{\beta}_3}} \sim t_{n-(3-1)} = t_{15}.$$

(Fortsettes på side 3.)

Et $(1-\alpha)100\%$ konfidensintervall finnes fra

$$P(-t_{\alpha/2,15} < \frac{\widehat{\beta}_3 - \beta_3}{S_{\widehat{\beta}_3}} < t_{\alpha/2,15}) = 1-\alpha.$$

Med $\alpha = 0.01$ trenger vi $t_{0.005,15} = 2.947$, slik at et 99% konfidensintervall for β_3 blir $\widehat{\beta}_3 \pm 2.947 * S_{\widehat{\beta}_3} = -3.96 \pm 2.947 \cdot 0.8437 = -3.96 \pm 2.4864 = (-6.45, -1.47)$.
Hvis vi velger ut 19 personer tilfeldig og beregner et slik 99% konfidensintervall for β_3 mange ganger, vil vi i det lange løp få intervaller som dekker den sanne verdien av β_3 i 99% av tilfellene.

f) "Multiple R-squared" finnes ved $R^2 = 1 - SSE/SST$, eller som kvadratet av korrelasjonen mellom observerte responser y_i og predikerte responser $\hat{y}_i$, $i = 1, ...n$. Tall mellom 0 og 1, som gir andelen av variasjonen i y som kan forklares av modellen. I plottet ser vi klart at det er forskjell i stigningstall for kvinner og menn. Bør nevne residualplott og normalfordelingsplott for residualene.

Oppgave 3.

a) Med modellen $y_i = \beta x_i + \epsilon_i$, $i = 1, ..., n$, setter vi opp utttrykket for kvadratavvikene $\sum_{i=1}^n (y_i - \beta x_i)^2$. Deriverer med hensyn på β og setter lik 0:

$$\frac{\partial}{\partial \beta} \sum_{i=1}^n (y_i - \beta x_i)^2 = 0$$

$$2 \sum_{i=1}^n (y_i - \beta x_i)(-x_i) = 0$$

$$\sum_{i=1}^n y_i x_i - \beta \sum_{i=1}^n x_i^2 = 0$$

$$\widehat{\beta}_{MK} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

$$E(\widehat{\beta}_{MK}) = \frac{\sum_{i=1}^n x_i E(y_i)}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i \beta x_i}{\sum_{i=1}^n x_i^2} = \beta$$

$$V(\widehat{\beta}_{MK}) = \frac{\sum_{i=1}^n x_i^2 V(y_i)}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sum_{i=1}^n x_i^2 \sigma^2}{(\sum_{i=1}^n x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

(Fortsettes på side 4.)

b)

$$E(\widehat{\beta}_A) = \frac{\sum_{i=1}^n E(y_i)}{\sum_{i=1}^n x_i} = \frac{\beta \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i} = \beta$$

$$V(\widehat{\beta}_A) = \frac{\sum_{i=1}^n V(y_i)}{(\sum_{i=1}^n x_i)^2} = \frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2} = \frac{\sigma^2}{n\bar{x}^2}$$

Skal sammenligne

$$\frac{\sigma^2}{n\bar{x}^2} \quad \text{og} \quad \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

Vet at $\sum_{i=1}^n (x_i - \bar{x})^2 \geq 0$ alltid er sant. Ved enkel regning finner vi fra dette at $\sum_{i=1}^n x_i^2 \geq n\bar{x}^2$ alltid. Derfor er

$$\frac{\sigma^2}{\sum_{i=1}^n x_i^2} \leq \frac{\sigma^2}{n\bar{x}^2},$$

dvs. $V(\widehat{\beta}_{MK}) \leq V(\widehat{\beta}_A)$ og vi velger derfor $\widehat{\beta}_{MK}$.

Oppgave 4.

a) Skal teste $H_0 : \lambda = 1$ mot $H_a : \lambda > 1$ med $\alpha = 0.05$. Naturlig å forkaste H_0 når $X \geq k$, der k finnes fra

$$P(X \geq k \mid \lambda = 1) \leq 0.05$$
$$1 - P(X < k \mid \lambda = 1) \leq 0.05$$
$$P(X < k \mid \lambda = 1) \geq 0.95$$

Fra Tabell: $P(X \leq 3 \mid \lambda = 1) = P(X < 4 \mid \lambda = 1) = 0.981$.
Så $k = 4$, dvs. man må minst observere 4 tilfeller for å kunne forkaste på nivå 0.05.

b) $P(X \geq 4 \mid \lambda = 2) = 1 - P(X \leq 3 \mid \lambda = 2) = 1 - 0.857 = 0.143$

$$P(\text{Type II feil}) = 0.857$$

c) Siden $X_i \sim \text{Poisson}(n_i p)$, har vi

$$P(X_i = x_i) = \frac{(n_i p)^{x_i} e^{-n_i p}}{x_i!}.$$

Likelihood for m år blir da

$$L(p) = \prod_{i=1}^m \frac{(n_i p)^{x_i} e^{-n_i p}}{x_i!}$$

(Fortsettes på side 5.)

Log-likelihood blir

$$loglik(p) = \sum_{i=1}^{m} [x_i \log(n_i p) - n_i p - \log(x_i!)]$$

$$\begin{aligned} \frac{\partial}{\partial p} loglik(p) &= \sum_{i=1}^{m} \left[\frac{x_i n_i}{n_i p} - n_i \right] \\ &= \sum_{i=1}^{m} \left[\frac{x_i}{p} - n_i \right] \\ &= \frac{\sum_{i=1}^{m} x_i}{p} - \sum_{i=1}^{m} n_i \end{aligned}$$

Setter lik 0 og finner MLE

$$\widehat{p} = \frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m} n_i}$$

$$E(\widehat{p}) = \frac{\sum_{i=1}^{m} E(X_i)}{\sum_{i=1}^{m} n_i} = \frac{\sum_{i=1}^{m} n_i p}{\sum_{i=1}^{m} n_i} = p$$

$$V(\widehat{p}) = \frac{\sum_{i=1}^{m} V(X_i)}{(\sum_{i=1}^{m} n_i)^2} = \frac{\sum_{i=1}^{m} n_i p}{(\sum_{i=1}^{m} n_i)^2} = \frac{p}{\sum_{i=1}^{m} n_i}$$