Løsningsforslag INT010 vår 2013

Oppgavene er laget av Jarle Møen

Sensorene anbefales å gi inntil 10 poeng på hvert delspørsmål. Sensorkorpset setter i felleskap karaktergrenser basert på prosentskår. Husk at dette er en åpen bok-eksamen. Man bør derfor ikke la seg imponere for mye av avskrift fra forelesningsnotatene. Selvstendig forståelse skal premieres.

Oppgave 1

a) Vi har én populasjon av OL-vinnere, så spørsmålet er om andelen født i første halvår er signifikant forskjellig fra 0,5. (Eller ekvivalent om andelen født i andre halvår er signifikant forskjellig fra 0,5.)

$Z = (\hat{p} - p)/\sqrt{p(1-p)/n} = \left(\frac{48}{85} - 0{,}50\right)/\sqrt{0{,}50(1-0{,}50)/85} = \underline{1{,}19}$

Z er tilnærmet standard normaltfordelt. Kritisk grense for tosidig test med 5 % signifikansnivå er 1,96. Vi kan altså ikke forkaste en hypotese om at OL-vinnerne er likt fordelt på første og andre halvår. (Kan og bruke ensidig test, men bør argumentere for det.)

[Forslag: 5 poeng for rett formel og 5 poeng for rett beregning og konklusjon.]

b) La det antall utøvere som bringer testobservatoren over signifikans være x. Vi skal da ha at $\left|\left(\frac{x}{85} - \frac{1}{12}\right)/\sqrt{\frac{1}{12}\left(1 - \frac{1}{12}\right)/85}\right| > 1{,}96.$

Løser vi ligningen får vi $x > \left(1{,}96 \cdot \sqrt{\frac{11}{144}/85} + \frac{1}{12}\right) \cdot 85 = 12{,}08$

eller $x < \left(-1{,}96 \cdot \sqrt{\frac{11}{144}/85} + \frac{1}{12}\right) \cdot 85 = 2{,}09.$

Vi må altså enten observere 13 eller flere, eller 2 eller færre, utøvere med fødselsmåned desember for å konkludere med signifikans.

[Forslag: 2 poeng trekk for bare å ha en grense, 1 poeng trekk for feil avrunding.]

c) Påstanden er at det vil være få utøvere født sent på året fordi de er yngst på det årskullet de konkurrerer i som barn og unge og derfor har mindre sjanse for å bli plukket ut som talenter. Vi kan argumentere for at vi på «teoretisk» grunnlag kan utelukke at det er spesielt gunstig å være født seint på året. H_A blir da $p<1/12$. (Uansett hvor høy andel utøvere vi observerer født i desember, vil vi ikke si at det taler mot nullhypotesen om $p=0{,}5$.)

d) Situasjonen tilsier kjikvadrattest for modelltilpasning. La p_i være sannsynligheten for at en tilfeldig nordmann er født i måned i. (Denne kan man beregne med utgangspunkt i SSB-statistikk for fødsler per måned for de aktuelle årskullene.) Testobservatoren blir $\chi 2 = \sum_{i=1}^{12} \frac{(f_i - e_i)^2}{e_i}$. Observerte frekvenser, f_i, er gitt i figur 1. Forventede frekvenser under H_0 om ingen forskjell mellom OL-vinnere og befolkningen forøvrig er gitt ved $e_i = 85 \cdot p_i$.

e) Vi skal nå teste om to andeler er lik hverandre. Testobservatoren er

$$Z = \left(\hat{p}_1 - \hat{p}_2 / \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}\right) = \left(\frac{16}{25} - \frac{15}{22} / \sqrt{\left(\frac{1}{25} + \frac{1}{22}\right)\frac{16+15}{25+22}\left(1 - \frac{16+15}{25+22}\right)}\right) =$$

$$\left(0{,}640 - 0{,}682 / \sqrt{0{,}0855 \cdot 0{,}660 \cdot 0{,}340}\right) = -0{,}042/0{,}139 = -0{,}302.$$

Kritisk grense for tosidig test på 5 %-nivå er ±1,96. Følgelig kan vi ikke forkaste en nullhypotese om like andeler.

[Forslag: 4 poeng for rett formelbruk. 3 poeng for rett utregning og 3 poeng for rett kritisk grense/konklusjon.]

Oppgave 2

a) Antall aviser i Oslo er 0,086·128=11.

b) Korrelasjonskoeffisienten mellom variablene opplag og no2 er -0,046. Det er altså en svakt negativ samanheng. (De er mindre enn nummer 1-aviser, men utkommer kun i større byer og er derfor større enn mange lokalaviser.)

c) Vi har følgende sammenhenger

$$\hat{\beta} = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\text{cov(Opplag, Redkost)}}{\text{var(Opplag)}} = \frac{Corr(\text{Opplag, Redkost}) \cdot SE(\text{Opplag}) \cdot SE(\text{Redkost})}{\text{var(Opplag)}}$$

$$= \frac{0.943 \cdot 27.549 \cdot 38.456}{38.456^2} = 0{,}676.$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = \overline{\text{Redkost}} - \hat{\beta} \cdot \overline{\text{Opplag}} = 14{,}875\text{-}0{,}676 \cdot 17{,}589 = 2{,}985.$$

[Forslag: 6 poeng for β (3 for formel, 3 for beregning), 4 poeng for α (2 for formel og 2 for beregning).]

d) Forventet redaksjonell kostnad blir 2,985+0,676·17,589=14,875 millioner kroner.

[Forslag: 7 poeng for riktig beregning, 3 poeng for riktig benevning.]

e) Et 95% konfidensintervall for E(Y|X) konstrueres som $\hat{Y} \pm t_{\alpha/2,n-2} \cdot S(\hat{Y})$ der der *t* er kritisk grense fra en *t*-fordeling. I vårt tilfelle blir det 1,979. Videre er

$$S(\hat{Y}) = S_e \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1)S_X^2}} = 9{,}184 \cdot \sqrt{\frac{1}{128} + 0} = 0{,}812.$$

Et 95 % konfidensintervall blir da14,875±1,979·0,812=14,875±1,607=[13,268;16,482].

[Forslag: 5 poeng for formler, 5 poeng for beregninger.]

f) Vi ser at når opplaget øker med 1 % øker også de redaksjonelle utgiftene med 1 %, alt annet likt. (*2p*) Videre ser vi at de redaksjonelle kostnadene øker med 14 % når antall utgivelser per uke øker med 1, alt annet likt. (*2p*) Disse sammenhengene har høy statistisk signifikans. (*2p*) Vi finner ikke signifikant effekt verken av interaksjonen mellom opplaget og antall utgivelser, om avisen er en no. 2-avis eller om den utkommer i Oslo. [Dette er litt overraskende.] (*2p*) Modellens prediksjonskraft er svært god, med en justert R^2 på 0,92. (*2p*).

[Poengforslag i parentes.]

g) En påstand om «eksponensiell vekst» må innebære at de redaksjonelle kostnadene vokser overproporsjonalt med opplaget. Vi ser ingen tegn til det i denne figuren. Tvert imot gjør de to største avisene at vi får en avtagende kurve.

h) Vi har at

$$\mathrm{Var}(\varepsilon/opplag) = \frac{1}{opplag^2}\mathrm{Var}(\varepsilon) = \frac{1}{opplag^2}[\mathrm{S}(\varepsilon)]^2 = \frac{1}{opplag^2}[opplag \cdot \sigma]^2$$

$$= \frac{opplag^2}{opplag^2} \cdot \sigma^2 = \sigma^2$$

Oppgave 3

a) Hvit støy må være det beste valget. Dersom prosessen var en tilfeldig gang ville verdien (antall plansjer på en forelesning) kunne vandre fritt og over tid bli arbitrært høyt eller lavt. Det er ikke mulig. Med hvit støy er antall plansjer en stokastisk variabel som trekkes fra en fordeling med konstant forventning og varians. Det kan være en rimelig tilnærming til hvor mange plansjer en rekker å gjennomgå på en totimers forelesning.

b) Forelesningsrekken består av ulike tema som strekker seg over flere forelesninger. I noen perioder er det stor vekt på teorigjennomgang og dermed mange plansjer. I andre temaer er det stor vekt på oppgavegjennomgang på tavlen og dermed mindre bruk av plansjer. Dette vil skape positiv autokorrelasjon. AR(1) kan være en enkel modell som fanger opp dette: $Y_t = \mu + \phi Y_{t-1} + u_t$ med $0 < \phi < 1$. En kilde til negativ autokorrelasjon kan være uforutsette avvik i forhold til planlagt gjennomgang på den enkelte forelesning. Går det fort har man bedre tid på neste forelesning. Kommer man på etterskudd må en gå litt fortere fram på neste forelesning. En mulig modell for en slik mekanisme er MA(1): $Y_t = \mu + u_t + \theta u_{t-1}$ med $-1 < \theta < 0$.

[Forslag: 5 poeng for en forklaring om hva som kan skape autokorrelasjon og 5 poeng for en matchende modell.]