Prelininära lösningar:
Tentamen
Tillämpad statistik A5 (15hp)
2016-05-31 uppdaterad

Statistiska institutionen, Uppsala universitet

Uppgift 1 (16 poäng)

A) (4p) Oddset för att som vuxen någon gång ha uppburit aktivitetsersättning är 1.16 gånger större för individer födda i december jämfört med individer födda i januari. Detta med hänsyn taget till eventuell betydelse av årskull.

B) (3p) Oddskvoten är kvoten mellan oddset för en händelse i en grupp dividerat med oddset för en händelse i en annan grupp. Om vi för enkelhetens skull ignorerar årskull kan detta skrivas som

$$OR = e^{\beta_1} = \frac{Odds(y = 1|x_1 = 1)}{Odds(y = 1|x_1 = 0)} = \frac{\frac{\Pr(y = 1|x_1 = 1)}{1 - \Pr(y = 1|x = 1)}}{\frac{\Pr(y = 1|x_1 = 0)}{1 - \Pr(y = 1|x = 0)}}.$$

Vi ser nu att om $\Pr(y = 1|x = 1)$ och $\Pr(y = 1|x = 0)$ är små (dvs om sannolikheten för en händelse är liten) så är $1 - \Pr(y = 1|x = 1)$ och $1 - \Pr(y = 1|x = 0)$ nära 1. Således är då

$$OR \approx \frac{\Pr(y = 1|x = 1)}{\Pr(y = 1|x = 0)} = RR$$

där RR är den relativa risken. Tolkningen av e^{β_1} blir nu (till skillnad från oddset som vi skrev i A): Sannolikheten att som vuxen någon gång ha uppburit aktivitetsersättning är 1.16 gånger större för individer födda i december jämfört med individer födda i januari. Detta med hänsyn taget till eventuell betydelse av årskull.

C) (4p)

- Mål: Undersök om födelsemånad är associerad med
- Parameter: O
- Hypoteser: $H_0 : OR = 1$ vs $H_1 : OR \neq 1$
- Förutsättningar: Stort stickprov. Många händelser.
- Beslutsregel: 5% signifikansnivå
- Beräkning: Det 95% konfidensintervallet 1.11-1.20 täcker inte 1.
- Eftersom konfidensintervallet inte täcker 1 så leder det till att vi, i detta fall, förkastar H_0.
- Svar: Vi kan, på 5% signifikansnivå, påvisa att födelsemånad är associerad med

D) (5p) Det är i princip inte nödvändigt att inkludera kön och föräldrars utbildning. Både kön och föräldrars utbildning förväntas vara orelaterat till födelsemånad, dvs förmodligen är barn till lågutbildade inte i större utsträckning är födda i december och förmodligen är andelen pojkar födda i december samma som andelen pojkar födda i januari. Grupperna förväntas vara lika i alla avseenden förutom födelsemånad.

Uppgift 2 (16 poäng)

Envägs variansanalys (observationsnummer har ingen betydelse).

Modell: $y_{ij} = \mu + \alpha_i + e_{ij}$ $\quad i = 1,2,3 \quad j = 1,2,3,4,5,6$

$H_0: \mu_1 = \mu_2 = \mu_3$ alt. $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$

H_1:Alla μ_i är ej lika $\quad H_1$:Alla α_i är ej lika med noll

signifikansnivå: $\alpha = 0{,}05$

testfunktion: $F_{obs} = \frac{MSA}{MSE}$

förutsättningar: $e_{ij} = NID(0, \sigma_e^2)$

beslutsregel: H_0 förkastas om $F_{obs} > F_{a-1,n-a;\alpha} = F_{2,15;0.05} = 3{,}68$

$$A = \sum_{i=1}^{3}\sum_{j=1}^{6} y_{ij}^2 = 40\,533$$

$$B = \frac{(\sum\sum y_{ij})^2}{18} = \frac{849^2}{18} = 40\,044{,}50$$

$$C = \sum_{i=1}^{3} \frac{(\sum y_{ij})^2}{6} = \frac{289^2 + 263^2 + 297^2}{6} = 40\,149{,}83$$

$SST = A - B = 40\,533 - 40\,044{,}50 = 488{,}50$

$SSE = A - C = 40\,533 - 40\,149{,}83 = 383{,}17$

$SSA = C - B = 40\,149{,}83 - 40\,044{,}50 = 105{,}33$

ANOVA-tabell

Variationsorsak	f.g.	SS	MS	F-test
Landsdelar	2	105,33	52,67	2,062
Error	15	383,17	25,54	
Totalt	17	488,5		

beslut: $F_{obs} = 2{,}062 < 3{,}68$ $\quad H_0$ förkastas inte

Testresultatet ger, på 5% signifikansnivå, inget belägg för att den genomsnittliga tiden varierar mellan landsdelarna.

Uppgift 3 (26 poäng)

A) (4p) De $n = 6$ datapunkterna (x, y) är $(2, 800)$, $(3, 800)$, $(4, 600)$, $(5, 700)$, $(6, 600)$, $(7, 500)$. Tillämpning av minstakvadratmetoden ger att regressionslinjens lutning är

$$\hat{\beta} = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1} x_i y_i - \bar{y}\bar{x}n}{\sum_{i=1} x_i^2 - \bar{x}^2 n} = \frac{17000 - 4.5 \cdot 666.6667 \cdot 6}{139 - 4.5^2 \cdot 6} = \frac{1000}{17.5} = -57.14.$$

Notera att $SS_{xy} = 1000$ och $SS_{xx} = 17.5$.

B) (4p) Interceptet ges av

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 666.6667 - (-57.14)4.5 = 923.8 \text{ tusentals kronor.}$$

C) (9p)

- Mål: Undersöka om det i branschen finns en linjär association mellan antal anställa (x) och personalkostnader per anställd (y).
- Modell: $y = \beta_0 + \beta_1 x + \varepsilon$
- Parameter: β_1
- Estimator: $\hat{\beta}_1$
- Hypoteser: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- Förutsättningar: (i)-(v) anses i uppgiften vara uppfyllda.
- Testfunktion: $t = (\hat{\beta}_1 - 0)/\sqrt{\hat{V}(\hat{\beta}_1)}$ som är t-fördelad med $n - (k + 1) = 6 - 2 = 4$ frihetsgrader om H_0 är sann.
- Beslutsregel: Tvåsidigt test med signifikansnivån 5%, dvs $\alpha = 0.05$. Förkasta H_0 om $|t_{obs}| > t_{krit}$, där $t_{krit} = t_{4,0.025} = 2.776$.
- Beräkning: Residualerna från den skattade regressionen i A) är -9.5238, 47.6190, -95.2381, 61.9048, 19.0476, -23.8095, vilket innebär att $SSE = 16190.46$ och $s_\varepsilon^2 = 16190.46/4 = 4047.6$ Från A) fick vi att $SS_{xx} = 17.5$, vilket ger att $V(\hat{\beta}_1)$ skattas med

  $$\hat{V}(\hat{\beta}_1) = s_\varepsilon^2 / SS_{xx} = 4047.6/17.5 = 231.29.$$

  Testfunktionens observerade värde blir

  $$t_{obs} = -57.14/\sqrt{231.29} = -3.76$$

- Beslut: Eftersom $|t_{obs}| = 3.76 > 2.776 = t_{krit}$ förkastar vi H_0.
- Svar: Vi kan på 5% signifikansnivå påvisa att i branschen så är antal anställa är associerat med personalkostnader per anställd.

D) (6p)

- Mål: Beräkna ett 90% prediktionsintervall för personalkostnad per anställd för ett företag med 6 anställda.

- Modell: $y = \beta_0 + \beta_1 x + \varepsilon$
- Estimator: $\hat{y}$
- Förutsättningar: (i)-(v) anses i uppgiften vara uppfyllda.
- Beräkning: Ett 90% prediktionsintervall ges av

$$\hat{y} \pm t_{n-(k+1),\alpha} \sqrt{\hat{V}(\hat{y}) + s_\varepsilon^2}.$$

Om $x = 6$ så är $\hat{y} = 580.952$, $t_{4,0.05} = 2.132$, $s_\varepsilon^2 = 4047.6$,

$$\hat{V}(\hat{y}) = s_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right) = 4047.6 \left(\frac{1}{6} + \frac{(6 - 4.5)^2}{17.5} \right) = 1195.$$

Insättning av värden ger intervallet

$$580.95 \pm 201.03$$

- Svar: För ett företag med 6 anställda så är med 95% sannolikhet personalkostnaden per anställd mellan 380 000 till 782 000 kronor.

E) (3p) Det blir tveksamt om normalfördelningsantagandet håller eftersom normalfördelningen är en fördelning för kontinuerliga variabler och vi nu, efter den grova avrundning, enbart har ett fåtal diskreta värden som möjliga utfall. Dessutom har vi få observationer så CGS gäller inte.

Uppgift 4 (22 poäng)

A) (4p) Vår uppfattning om $p = 0.3$ och vi drar ett OSU-MÅ. Eftersom konfidensintervallet för en andel är $2 \times felmarginalen$ så är precisionskravet är

$$Precisionkrav \geq 2 \cdot z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Insättning av värden ger

$$0.02 \geq 2 \cdot 1.96\sqrt{\frac{0.3(1-0.3)}{n}}$$

Vi löser ut n

$$0.02^2 \geq 2^2 \cdot 1.96^2 \frac{0.3(1-0.3)}{n}$$

$$n \geq 2^2 \cdot 1.96^2 \frac{0.3(1-0.3)}{0.02^2}$$

$$n \geq 8067.36$$

vilket avrundas till att vi måste ha en urvalsstorlek på 8068 för att uppnå precisionskravet. Rätt ges även för ett svar som baserat på att iterativt testa för olika värden för n.

B) (4p) Nu har vi en ändlig population, $N = 1500$ och drar ett OSU-UÅ. Eftersom konfidensintervallet för en andel är $2 \times felmarginalen$ så är precisionskravet är

$$Precisionkrav \geq 2 \cdot z_{\alpha/2}\sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}}$$

Insättning av värden ger

$$0.02 \geq 2 \cdot 1.96\sqrt{\left(\frac{1500-n}{1500-1}\right)\frac{0.3(1-0.3)}{n}}$$

och löser vi ut n får vi att

$$n \geq \frac{1500}{\frac{1499 \cdot 0.02^2}{0.3 \cdot (1-0.3) \cdot 2^2 \cdot 1.96^2} + 1} = 1264.96$$

vilket avrundas till att vi måste ha en urvalsstorlek på 1265 för att uppnå precisionskravet. Rätt ges även för ett svar som baserat på att iterativt testa för olika värden för n.

A) (14p) Vi börjar med att konstatera att de 100 elefanternas totala vikt, τ, är en parameter och ej slumpmässig. Antingen väger elefanterna mindre än 540 ton eller så väger de 540 ton eller mer. Det innebär att sannolikheten att båten klarar att skeppa över alla elefanter utan att sjunka är

$\Pr(\tau > 540) = 0$ eller $\Pr(\tau > 540) = 1$ beroende på värdet på τ (som är okänt men ej slumpmässigt).

Denna kunskap hjälper emellertid inte skogsägaren i sitt beslutsfattande (chansa eller inte chansa). Därför vill vi nu med ett urval göra så att skogsägaren kan få en uppfattning om värdet på parametern. För att göra detta genomför vi en hypotesprövning (alternativt beräknar ett konfidensintervall). Skogsägaren nöjer sig (kanske förvånansvärt nog) med signifikansnivån 5% och mothypotesen är naturligtvis ensidig. Hypoteserna formuleras

- $H_0 : \tau \geq 540$ vs $H_1 : \tau < 540$.

Populationen består av två strata med $N_1 = 60$ elefantjurar och $N_2 = 40$ elefantkor. Stickprovsstorlekarna är $n_1 = 6$ och $n_2 = 4$. Från urvalet erhålls stickprovsmedelvärdena $\bar{x}_1 = 6.2$ och $\bar{x}_2 = 4.1$ och stickprovsvarianserna $s_1^2 = 4$ och $s_2^2 = 2.25$. Vi börjar med att skatta elefanternas totala vikt:

$$\hat{\tau}_{st} = N_1\bar{x}_1 + N_2\bar{x}_2 = 536$$

Vi ser att $\hat{\tau}_{st} = 536 < 540$. Skattningen är väntevärdesriktig pga OSU-UÅ och indikerar att skogsägaren kanske kan chansa. Emellertid ger punktskattnigen ingen information om precisionen som skattning har. Eftersom urvalen från respektive stratum är obereonde av varandra samt att vi har OSU-UÅ skattar vi därför variansen för $\hat{\tau}$ med

$$\begin{aligned}\hat{V}(\hat{\tau}) &= N_1^2\left(1 - \frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + N_2^2\left(1 - \frac{n_2}{N_2}\right)\frac{s_2^2}{n_2} \\ &= 60^2\left(1 - \frac{6}{60}\right)\frac{4}{6} + 40^2\left(1 - \frac{4}{40}\right)\frac{2.25}{4} = 2970.\end{aligned}$$

Stickprovet är litet vilket innebär att ett antagande om att elefanternas vikt är normalfördelad måste göras (vilket kan vara ett rimligt vad gäller en variabel som vikt). Dessutom så är $V(\hat{\tau})$ okänd. Under dessa förutsättningar är testfunktionen

$$t = \frac{\hat{\tau}_{st} - \tau_0}{\sqrt{\hat{V}(\hat{\tau}_{st})}}$$

t-fördelad med $n - 1$ frihetsgrader om H_0 är sann. Eftersom $\alpha = 0.05$, $n - 1 = 9$ förkastas nollhypotesen om $t_{obs} > t_{krit} = t_{9,0.05} = 1.833$. Vi får att

$$t_{obs} = \frac{536 - 541}{\sqrt{2970}} = -0.074$$

och eftersom $|t_{obs}| = 0.074 < 1.833 = t_{krit}$ kan vi inte förkasta H_0. Eftersom vi på 5% procents signifikansnivå inte kan påvisa att den totala vikten är mindre än 541 ton bör skogsägaren definitivt inte chansa.

Uppgift 5 (20 poäng)

A) (8p) Gör prognoser:

$$\hat{y}_{120+1} = 4343,2 + 3,246 \times (120+1) + 21,7 = 4757,71$$
$$\hat{y}_{120+2} = 4343,2 + 3,246 \times (120+2) + 36,1 = 4775,35$$
$$\hat{y}_{120+3} = 4343,2 + 3,246 \times (120+3) + 62,7 = 4805,18$$
$$\hat{y}_{120+4} = 4343,2 + 3,246 \times (120+4) + 186,4 = 4932,14$$

B) (6p) En linjär trend likt vad vi har ovan kan fungera bra och vara rimlig under en kortare tidsperiod. Däremot är det ofta orimligt att anta att det kommer fortsätta på samma sätt i framtiden, vilket gör att man bör vara försiktig vad gäller längre prognoser. Hade vi haft en längre tidsserie är det troligt att vi inte kunnat använda oss av en linjär trend.

C) (6p) I det här fallet, då serien uppvisar både en trend och säsongsmässig variation, så är trippel exponentiell utjämning (Holt-Winters metod med trend och säsong) lämplig att använda. Eftersom vi i denna metod hela tiden uppdaterar våra nivå-, trend- och säsongsskattningar så anpassar metoden sig utifall att t ex den relativt linjära trendutvecklingen i figuren skulle brytas.