Tentamen
Tillämpad statistik A5 (15hp)
2016-05-31

Statistiska institutionen, Uppsala universitet

Upplysningar

1. Tillåtna hjälpmedel: Miniräknare, A4/A8 Tabell- och formelsamling (alternativ Statistik för samhällsplanerare Tabell- och formelsamling) samt nuvarande formelsamling för A5. Formelsamlingar för A4/A8 samt A5 som användes HT2014-VT2015 är också tillåtna. **Inga anteckningar är tillåtna i formelsamlingarna.**
2. Skrivtid: **8.00-13.00**. Skrivningen omfattar 5 uppgifter, om sammanlagt 100 poäng.
3. För varje uppgift anges den maximala poäng som kan erhållas. Om en uppgift är uppdelad på deluppgifter anges den maximala poängen för varje deluppgift. Ibland kan inte deluppgifterna bedömas oberoende av varandra, vilket kan innebära att poäng inte utdelas på en senare uppgift om inte tidigare deluppgift lösts på ett i princip riktigt sätt. Dock gäller att utdelad poäng för varje deluppgift aldrig kan vara negativ.
4. Om du känner dig osäker på någonting (skrivningens genomförande, någon formulering i en uppgift, om något hjälpmedel är otillåtet), fråga då jourhavande skrivningsvakt eller den skrivningsansvariga läraren (besök, alternativt telefon).
5. Efter skrivningens slut får du behålla sidorna med frågeställningarna. Preliminära lösningar anslås på Studentportalen.

Uppmaningar

1. Följ noga de anvisningar som finns på skrivningsförsättsbladet.
2. Alla lösningar ska redovisas i en form som gör det lätt att följa din tankegång! Motivera alla väsentliga steg i lösningen. Ange alla antaganden du gör och alla förutsättningar du utnyttjar. Alla uppgifter kräver en verbal slutsats.
3. Vid konfidensintervall måste du ange vad intervallet avser att täcka samt teckna intervallet i symbolform innan de numeriska uppgifterna insätts.
4. Vid alla hypotestest måste du ange H_0, H_1, signifikansnivå, testfunktion, frihetsgrader, förkastelseområde och resultat.
5. Vid variansanalys måste du utöver vad som nämns ovan ange modell.

Uppgift 1 (16 poäng)

Försäkringskassans rapport *Barns relativa ålder och funktionsnedsättning*[1] syftade till att studera betydelsen av att vara född i slutet av året för aktivitetsersättning (tidigare kallat förtidspension) länge fram i livet. Rapporten inkluderade årskullar med alla individer födda mellan 1974 och 1994 och jämförde individer födda i december i en årskull med individer födda i januari året därefter. En modell som skattades i rapporten var

$$\ln\left(\frac{\Pr(y=1|x_1,x_2,\ldots,x_k)}{1-\Pr(y=1|x_1,x_2,\ldots,x_k)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$$

där $y=1$ om individen innehaft aktivitetsersättning efter 19 års ålder, $x_1=1$ om individen är född i december och $x_1=0$ om individen är född i januari och x_2-x_k är dummy-variabler för årskullarna 1975-1994 med 1974 som referens. Ett resultat sammanfattades så här i text:

> *Oddskvoten för barn födda i december att ha uppburit någon form av förtidspension var 1,16 (1,11–1,20).*

A) (4p) Ge en verbal tolkning av skattningen för e^{β_1}.

B) (3p) I rapporten anges att cirka 4% någon gång har erhållit aktivitetsersättning, så aktivitetsersättning är att betrakta som relativt sällsynt. Påverkar detta hur du kan tolka e^{β_1}? Motivera!

C) (4p) Genomför en hypotesprövning som på 5% signifikansnivå undersöker om födelsemånad är associerad med aktivitetsersättning någon gång efter 19 års ålder. Du behöver här inte ange testfunktion, frihetsgrader eller förkastelseområde utan utnyttja istället att parentesen i citatet ovan anger ett 95% konfidensintervall för oddskvoten.

D) (5p) Författarna har även tillgång till variabeln kön samt variabler som beskriver föräldrarnas utbildningsnivå. Med utgångspunkt från undersökningens syfte, är det nödvändigt att inkludera dessa variabler i modellen ovan? Ge en utförlig motivering!

[1] Socialförsäkringsrapport 2016:3, Försäkringskassan. (Notera att formuleringar kan vara ändrade från originalrapporten.)

Uppgift 2 (16 poäng)

Den stora bankkoncernen Omega bedriver bland annat privatekonomisk rådgivning i sina bankkontor. Ledningen önskar undersöka om den genomsnittliga tiden som åtgår vid varje rådgivningstillfälle varierar mellan tre olika landsdelar. I tabellen nedan redovisas tidsåtgången i minuter för 6 olika tillfällen valda slumpmässigt bland det stora antal rådgivningstillfällen som ägt rum i respektive landsdel under april månad. Summor som kan vara användbara redovisas också.

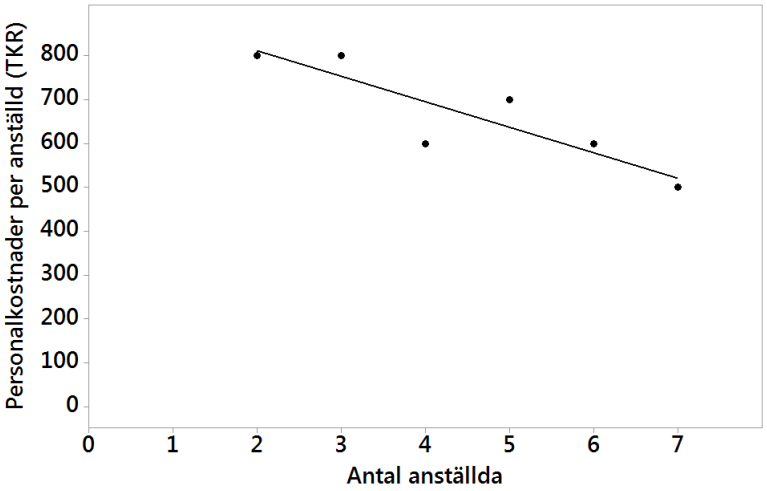
	Landsdel			
Observation nummer	1	2	3	Summa
1	49	53	53	155
2	43	41	60	144
3	47	44	46	137
4	54	40	49	143
5	50	44	42	136
6	46	41	47	134
Summa	289	263	297	849

Summan av alla kvadrerade observationer är 40 533.

Undersök med hjälp av en hypotesprövning på 5% signifikansnivå om den genomsnittliga tiden varierar mellan de tre landsdelarna.

Uppgift 3 (26 poäng)

Du har från en branschorganistation erhållit nedanstående figur. Figuren baseras på ett slumpmässigt urval av sex företag från en viss bransch. Branschen består av ett mycket stort antal företag. På x-axeln anges antal anställda och på y-axeln anges personalkostnad per anställd i tusentals kronor (avrundad till närmsta hundratusen kronor).



A) (4p) Regressionslinjen i figuren har erhållits med minstakvadratmetoden. Använd figurens datapunkter för att beräkna regressionslinjens lutning.

B) (4p) Vid vilket värde på y-axeln skär regressionslinjen y-axeln?

Du vill inte enbart studera regressionslinjens lutning i stickprovet, utan målet är att dra slutsatser om det generellt för branschen finns ett samband mellan antal antällda och personalkostnader per anställd. Vi ställer därför upp modellen

$$\text{Personalkostnad per anställd i tusentals kronor} = \beta_0 + \beta_1 \text{Antal anställda} + \varepsilon$$

Anta att alla nödvändiga förutsättningar gäller.

C) (9p) Genomför ett hypotestest för att på 5% signifikansnivå undersöka om det i branschen finns ett samband mellan antal anställda och personalkostnader per anställd.

D) (6p) Beräkna ett 90% prediktionsintervall för personalkostnad per anställd om vi slumpmässigt skulle välja ett företag med 6 anställda. Tolka intervallet!

E) (3p) Som utfall används data som är avrundad till hela hundratusen kronor. För vilken förutsättning är det tveksamt om den är uppfylld i och med detta? Motivera!

Uppgift 4 (22 poäng)

Vi är, för ett visst företag, intresserade av att skatta andelen män, p, som har sjukfrånvaro pga arbetsmiljön. Företaget har 1500 anställda. Vi vet att tre av tio anställda i ett annat företag i samma bransch har sjukfrånvaro pga arbetsmiljön. För att undersöka detta väljer vi en urvalsundersökning.

A) (4p) Om vi väljer OSU med återläggning som urvalsmetod, vilken urvalsstorlek ska väljas så är längden på ett konfidensintervall med 95% konfidensgrad är mindre än 0.02?

B) (4p) Om vi väljer OSU utan återläggning som urvalsmetod, vilken urvalsstorlek ska väljas så är längden på ett konfidensintervall med 95% konfidensgrad är mindre än 0.02?

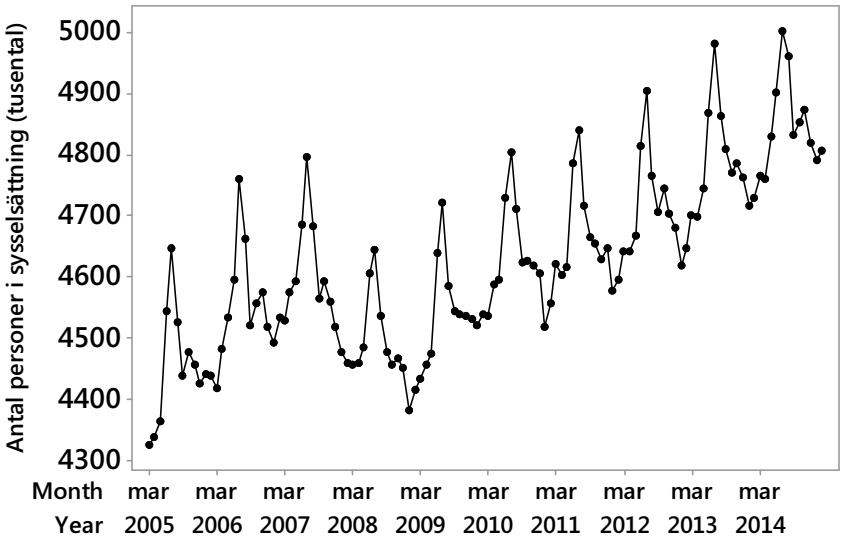
En skogsägare i Thailand har 100 arbetselefanter (60 elefanttjurar, 40 elefantkor). Skogsägaren vill nu transportera elefanterna till ett nytt område. Problemet är att elefanterna måste skeppas över en djup flod. Skogsägaren har dock tillgång till en båt som klarar av 540 tons last innan den sjunker. Nu ställer sig skogsägaren (den inte särskilt förvånande) frågan: "Hur ska det gå om alla elefanter lastas på båten och skeppas samtidigt?"

Skogsägaren måste därför få en uppfattning om elefanternas totala vikt, men eftersom det inte är enkelt att väga elefanter genomför skogsägaren ett proportionellt stratiferat urval och väger 6 elefanttjurar och 4 elefantkor. Medelvikten bland elefanttjurarna är 6.2 ton och medelvikten bland elefantkorna är 4.1 ton. Vidare är stickprovsvariansen för elefanttjurarna 4 och stickprovsvariansen för elefantkorna är 2.25.

C) (14p) Vad är sannolikheten att båten klarar att skeppa över alla elefanter utan sjunka? Bör skogsägaren chansa? Var noga med att ange förutsättningar samt ta hänsyn till urvalsdesignen i dina beräkningar.

Uppgift 5 (20 poäng)

Arbetskraftsundersökningarna (AKU) är en undersökning som beskriver utvecklingen på arbetsmarknaden för Sveriges befolkning i åldern 15–74 år. Undersökningen genomförs varje månad och är grunden för den officiella arbetslöshets- och sysselsättningsstatistiken.



Figur 1: Antal sysselssatta personer (15-74 år). Källa: SCB.

I figur 1 visas antalet sysselsatta personer (i tusental) under tidsperioden 2005:03-2016:02 (dvs från mars 2005 till februari 2016). För tidsserien har en regression med linjär trend och dummyvariabler för säsongsvariationen anpassats enligt:

$$y_t = \beta_0 + \beta_1 t + \beta_2 M_{2,t} + \beta_3 M_{3,t} + \beta_4 M_{4,t} + \beta_5 M_{5,t} + \beta_6 M_{6,t} + \beta_7 M_{7,t} + \beta_8 M_{8,t} + \beta_9 M_{9,t} + \beta_{10} M_{10,t} + \beta_{11} M_{11,t} + \beta_{12} M_{12,t} + \epsilon_t, \quad t = 1, 2, \ldots, 120.$$

där y_t = Antal sysselsatta personer i tusental, ϵ_t är en felterm och $M_{j,t} = 1$ om tidpunkten t tillhör månad j och $M_{j,t} = 0$ annars. Utskriften från skattningen av modellen i Minitab finns till din hjälp nedan.

```
Regression Analysis: Sysselsättning versus t; M2; M3; M4; M5; M6; M7;
M8; M9; M10; M11; M12

Model Summary

S    R-sq  R-sq(adj)  R-sq(pred)
56,1141  86,71%     85,22%      83,32%
```

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4343,2	20,2	214,97	0,000	
t	3,246	0,149	21,84	0,000	1,01
M2	15,5	25,1	0,62	0,539	1,83
M3	21,7	25,1	0,86	0,390	1,84
M4	36,1	25,1	1,44	0,154	1,84
M5	62,7	25,1	2,49	0,014	1,84
M6	186,4	25,1	7,42	0,000	1,84
M7	276,2	25,1	11,00	0,000	1,84
M8	163,8	25,1	6,52	0,000	1,83
M9	77,7	25,1	3,10	0,003	1,83
M10	83,6	25,1	3,33	0,001	1,83
M11	73,3	25,1	2,92	0,004	1,83
M12	45,8	25,1	1,83	0,071	1,83

A) (8p) Gör prognoser för perioden 2016:03-2016:06, dvs 1-4 månader framåt i tiden.

B) (6p) Diskutera rimligheten i att anta en linjär trend. På vilket sätt kan detta vara riskfyllt?

C) (6p) Ge exempel på en annan prognosmetod som skulle vara lämplig i det här fallet. Motivera tydligt varför du tycker att denna prognosmetod är ett bra alternativ.