STOCKHOLMS UNIVERSITET VT2023
Statistiska institutionen
Mattias Villani

SDAI (ST1101), Tentamen 2, 6 hp, första tentamenstillfället

Kurs: Statistik och dataanalys I, 15 hp
Tentamensdatum: 2023-03-20

Skrivtid:	kl. 8 - 13 (5 timmar)
Godkända hjälpmedel:	Miniräknare utan lagrade formler och text
Bifogade hjälpmedel:	Formel- och tabellsamling för Statistik och dataanalys I, 15 hp

Tentamen består av 5 uppgifter, uppdelade i deluppgifter.
Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas.
- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.
- Kontrollera alltid dina beräkningar och lösningar! Slarvfel kan också ge poängavdrag!
- Om du inte lyckas lösa en deluppgift och behöver det svaret för en senare deluppgift så kan du hitta på värdet för att kunna göra beräkningarna i de efterföljande uppgifterna.
- I beräkningar från R-utskrifter får du utgå från det som är givet.

Tentamen kan maximalt ge 100 poäng och för godkänt resultat krävs minst 50.

Betygsgränser:
A: 90–100p
B: 80–89p
C: 70–79p
D: 60–69p
E: 50–59p
Fx: 40–49p
F: 0 – 40p

OBS! Fx och F är underkända betyg som kräver omexamination.
Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

Lösningsförslag läggs ut på Athena efter tentamen i samband med rättningen.

Lycka till!

Uppgift 1 (16 poäng)

Låt A och B vara två oberoende händelser med $P(A) = 0.3$ och $P(B) = 0.2$.

(a) Vad är sannolikheten att både A och B inträffar? (4 p)

(b) Vad är sannolikheten att åtminstone någon av de två händelserna inträffar? (4 p)

(c) Vad är sannolikheten att ingen av de två händelserna inträffar? (4 p)

(d) Vad är sannolikheten för A givet att B har inträffat? (4 p)

Uppgift 2 (20 poäng) Sannolikheten att du kommer i tid till föreläsningen är 0.8. Antag att samma sannolikhet gäller för alla vardagar, och att förseningar vid olika dagar är oberoende av varandra.

(a) Vad är sannolikheten att du under nästa vecka blir försenad enbart på måndag och onsdag? (4 p)

(b) Vad är sannolikheten att din första försening för veckan blir först på fredag? (4 p)

(c) Vad är sannolikheten att du under nästa vecka blir försenad tre av fem vardagar? (6 p)

(d) I ett försök att skärpa dig har du lovat din kompis att betala henne 100 kr för varje försening under veckans fem arbetsdagar. Beräkna din förväntade kostnad för detta under den kommande arbetsveckan, givet att du inte ändrar ditt beteende. (6 p)

Uppgift 3 (20 poäng)

Ett stort globalt företag har gjort en undersökning bland 230 av deras kunder där de bland annat ställde frågan: 'Skulle du rekommendera vår service till andra personer?'. 172 personser svarade 'ja' , resterande svarade 'nej', eller 'vet inte'.

(a) Låt p vara andelen i populationen som skulle rekommendera servicen åt andra, och $\hat{p}$ andelen i stickprovet. Beräkna en normalapproximation av samplingfördelningen för $\hat{p}$. Ange eventuella antaganden för approximationen, och undersök dessa antaganden där det är möjligt. (11 p)

(b) Gör ett 95% konfidensintervall för populationsandelen p. Tolka intervallet på ett sätt som *visar* att du *förstår* innebörden med ett konfidensintervall. (9 p)

Uppgift 4 (22 poäng)

På ett lokalt gym tränar man i genomsnitt 54.3 minuter per gympass. Gymmet har fått klagomål att de spelar stressande musik i högtalarna. I ett försök att göra kunderna nöjda byter de ut alla spellistor med förhoppning att personerna ska trivas bättre och gymma längre per pass.

(a) De väljer slumpmässigt ut 5 olika kunder och observerar hur länge de gymmar efter att spellistorna har bytts ut. Medelvärdet av dessa 5 personers gympass efter byte av spellistorna var 66.8 minuter och standardavvikelsen var 22.26 minuter. Utför ett hypotestest på 5% signifikansnivå för att undersöka om medelträningstiden för gymmets kunder har förändrats efter byte av spellistorna. Ställ upp hypoteser, utför testet och dra korrekta slutsatser. Vilka antaganden måste du göra för att utföra testet? (9 p)

(b) En extrajobbare i personalen har läst lite statistik och ifrågasätter att gymmet verkar ta det tidigare genomsnittet på 54.3 minuter per gympass som känt. Hon menar att det genomsnittet på 54.3 minuter bara är en (osäker) skattning. Hon föreslår att man istället jämför med träningstiderna för de 5 utvalda kunderna innan man bytte spelistorna. Data för de fem personernas sista träningspass innan bytet och från första passet efter bytet ges i tabellen nedan. Ställ upp och utför ett hypotestest som testar om bytet av spellistor har påverkat längden på passen för gymmets kunder. Vilka antaganden måste du göra för att utföra testet? (13 p)

Person:	1	2	3	4	5	Medelvärde	Standardavvikelse
Träningstid innan byte:	60	39	46	59	33	47.40	11.97
Träningstid efter byte:	72	49	78	95	40	66.80	22.26

Uppgift 5 (22 poäng)

I den här uppgiften analyseras löner för 397 amerikanska universitetsprofessorer. Lönerna är månadslöner i motsvarande tusentals svenska kronor. Så `salary = 50` är en månadslön på 50000 kr.

(a) Vi börjar med analysera denna enkla linjära regression:

$$\texttt{salary} = \beta_0 + \beta_1 \cdot \texttt{yrs.since.phd} + \varepsilon, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon)$$

där `yrs.since.phd` är antalet år sedan doktorsexamen (med medelvärde 22.31 i stickprovet). Utskriften nedan visar den skattade modellen. Beräkna ett 95% prediktionsintervall för `salary` för en professor som fick doktorsexamen för 23 år sedan.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.312      2.014   41.37  < 2e-16 ***
yrs.service    0.650      0.092    7.06  7.5e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.8 on 395 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.11
```

(8 p)

(b) Vi lägger nu till dummyvariabeln `sexMale` i modellen:

$$\texttt{salary} = \beta_0 + \beta_1 \cdot \texttt{yrs.since.phd} + \beta_2 \cdot \texttt{sexMale} + \varepsilon, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon)$$

Variabeln `sexMale` är 1 om professorn är man och 0 om kvinna. Utskriften nedan ger den skattade modellen, men viss information i den vanliga utskriften i R har tagits bort.

```
Coefficients:
             Estimate Std. Error
(Intercept)   70.9848     3.9569
yrs.since.phd  0.7984     0.0903
sexMale        6.6030     3.9034
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.9 on 394 degrees of freedom
Multiple R-squared:  0.182,     Adjusted R-squared:  0.178
```

Testa om `sexMale` är en signifikant förklarande variabel på 5% signifikansnivå. Ställ upp hypoteser, teststatistikan med fördelningen under H_0 och utför testet. Dra slutsats. (7 p)

(c) Beräkna p-värdet för testet i Uppgift 5) och tolka det. Du får använda approximationer, om du kan motivera dem. (5 p)

(d) Vilken av de två modellerna är att föredra? Motivera. (2 p)

Lycka till!