STOCKHOLMS UNIVERSITET
Statistiska institutionen
Mattias Villani

HT2023

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

Kurs: Statistik och dataanalys I, 15 hp
Tentamensdatum: 2023-10-26

Skrivtid:	kl. 14 - 19 (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 händelser med sannolikheter $P(A) = 0.2$ och $P(B) = 0.5$. Låt $P(B|A) = 0.6$.

(a) Är A och B oberoende händelser? (4 p)

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

(c) Vad är sannolikheten att åtminstone någon av A och B inträffar? (4 p)

(d) Antag att vi nu vet att B har inträffat. Vad är nu sannolikheten för A? (4 p)

Uppgift 2 (22 poäng)

En ny internetsajt låter användare lägga upp videos till en engångskostnad av 50 kr per video. När en person (kreatör) lägger upp en ny video får den X antal visningar (views) under den första dagen och Y antal visningar under den andra dagen. Antag att $X \sim \text{Pois}(10)$ och $Y \sim \text{Pois}(5)$.

(a) Vad är det förväntade antal visningar under de två första dagarna? Ange eventuella antaganden du måste göra för beräkningen. (4 p)

(b) Vad är standardavvikelsen för antalet visningar under de två första dagarna? Ange eventuella antaganden du måste göra för beräkningen, och resonera om antagandena verkar rimliga. (5 p)

(c) Sajten ger kreatören 2 kronor per visning i reklamintäkter. Vad är den förväntade vinsten/förlusten efter de två första dagarna? (4 p)

(d) Skriv R-kommandot för att beräkna sannolikheten att kreatören gör en (positiv) vinst redan efter den första dagen? (4 p)

(e) Kreatören noterar att hen behövde vänta en hel timme innan första visningen kom. Antag att väntetiderna mellan visningar är exponentialfördelad. Är det onaturligt lång väntetid? Motivera ditt svar med en beräkning. (5 p)

Uppgift 3 (24 poäng)

Ett datorspel erbjuder s k *lådöppning* där man betalar 10 kr för en chans att vinna en speciell karaktär som sedan kan användas i spelet. Man kan köpa hur många lådöppningar som helst, spelet har obegränsat många karaktärer av varje sort. Lådöppningarna sker helt slumpmässigt och kan därför antas vara oberoende av varandra.

(a) Sannolikheten att få karaktären OverlordX i en lådöppning är 0.15, och anses vara känd. Vad är sannolikheten att få OverlordX först på 4:e lådöppningsförsöket? (4 p)

(b) Spelet erbjuder också spelare att köpa ett paket med fem öppningschanser för totalt 40 kr. Vad är sannolikheten att få karaktären OverlordX mer än en gång om du köper paketet? (5 p)

(c) Spelet har nyligen lanserat en ny karaktär UnderdogY, men det är inte känt vad sannolikheten är för att få den karaktären. En YouTube-kanal bestämmer sig för att ta reda på det, och köper 200 lådöppningar. De fick UnderdogY i 42 av de 200 lådöppningarna. Skapa ett 95%-igt konfidensintervall för sannolikheten att få UnderdogY. Var noga med att ange och kontrollera rimligheten i de eventuella antaganden som krävs för dina beräkningar. (7 p)

(d) Testa på 1% signifikansnivå om sannolikheten att få UnderdogY skiljer från sannolikheten att få OverlordX genom att beräkna testets p-värde. Ställ upp hypoteser och dra slutsatser. (8 p)

Uppgift 4 (15 poäng)

För att undersöka mängden zink i dricksvattnet gjordes mätningar på sex olika ställen. Vid varje mätplats mätte man mängden zink både vid botten och ytan. Mätningar ges i tabellen nedan.

	Mätplats nummer					
	1	2	3	4	5	6
Botten	4.30	2.66	5.67	5.31	7.07	7.16
Ytan	4.15	2.38	3.90	4.10	6.05	6.09

(a) Antag att mätningarna vid Botten vid de 6 olika platserna är oberoende och normalfördelade med väntevärde μ och standardavvikelse σ. Gör ett 95% konfidensintervall för μ. Vad menar vi mer exakt när vi säger den lite slarviga frasen att 'intervallet har 95% säkerhet'? (7 p)

(b) Antag även att mätningarna på Ytan är oberoende vid de 6 olika platserna och normalfördelade. Testa på 5% signifikansnivå om det är någon skillnad på den genomsnittliga mängden zink mellan Botten och Ytan i populationen. (8 p)

Uppgift 5 (23 poäng)

Ett lokalt glassföretag har introducerat smaken *Salty Brownie* i sortimentet inför den kommande sommaren, men den har sålt dåligt på försäsongen (våren), och går med förlust. Ägaren Mona har noterat att just den smaken inte verkar öka i försäljning under lite varmare dagar, vilket glassförsäljningen i allmänhet gör. Mona undersöker saken genom att samla in försäljningssiffror (`antalglass`) och temperatur (`temp`) under $n = 37$ dagar. Nedan ges en sammanfattning av data och en utskrift på anpassningen av den linjära regressionsmodellen (men vissa värden har tagits bort från utskriften)

$$\texttt{antalglass} = \beta_0 + \beta_1 \cdot \texttt{temp} + \varepsilon, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon).$$

```
     const         temp            helg           antalglass
 Min.   :1   Min.   :10.00   Min.   :0.0000   Min.   : 1.000
 1st Qu.:1   1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 6.000
 Median :1   Median :16.00   Median :0.0000   Median : 9.000
 Mean   :1   Mean   :15.46   Mean   :0.2703   Mean   : 8.973
 3rd Qu.:1   3rd Qu.:18.00   3rd Qu.:1.0000   3rd Qu.:12.000
 Max.   :1   Max.   :20.00   Max.   :1.0000   Max.   :19.000


Call:
lm(formula = antalglass ~ temp, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7904 -1.5985  0.2096  2.0177  9.0177

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8308     3.3961   0.834   0.4102
temp          0.3973     0.2160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.761 on 35 degrees of freedom
Multiple R-squared:  0.08814,   Adjusted R-squared:  0.06209
F-statistic: 3.383 on 1 and 35 DF,  p-value: 0.07436
```

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

(b) Gör en prognos med 95%-igt prediktionsintervall för antalet sålda *Salty Brownie* vid en dag med en temperatur på 30 grader. (7 p.)

(c) Mona vet sedan tidigare att hon säljer fler glassar under helger. Hon utökar därför modellen med en dummyvariabel `helg` som antar värdet 0 på en vardag och 1 på en helg:

$$\texttt{antalglass} = \beta_0 + \beta_1 \cdot \texttt{temp} + \beta_2 \cdot \texttt{helg} + \varepsilon, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon).$$

Utskriften nedan ger resultatet från en anpassning av denna multipla regressionsmodell. Tolka skattningen av β_2 och beräkna ett 90%-igt konfidensintervall för β_2. (5 p)

```
Call:
lm(formula = antalglass ~ temp + helg, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8255 -1.3956 -0.0031  1.9888  6.4187 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   1.9377     3.1198   0.621  0.53868  
temp          0.3925     0.1974                   
helg          3.5782     1.2725                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(d) Ange de fyra grundantaganden i populationsmodellen för linjär regression. (4 p)

Lycka till!