STOCKHOLMS UNIVERSITET
Statistiska institutionen
Mattias Villani

HT2023

# SDAI (ST1101), Tentamen 2, 6 hp, andra tentamenstillfället

**Kurs: Statistik och dataanalys I, 15 hp**
**Tentamensdatum: 2023-12-05**

| | |
|---|---|
| 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** (20 poäng)

Låt $A$ och $B$ vara två händelser. Vi har sannolikheterna $P(A) = 0.3$, $P(B^c) = 0.8$ och $P(B|A) = 0.5$, där $B^c$ är komplementshändelsen till $B$.

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

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

(c) Vad är sannolikheten att ingen av $A$ och $B$ inträffar? (4 p)

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

(e) Vad är sannolikhet att *exakt en* av $A$ eller $B$ inträffar? (4 p)

**Uppgift 2** (22 poäng)

En Youtube-kanal lägger ut videos om husdjur. Sannolikheten att en godtycklig användare gillar en video med katter (dvs ger en s k *like*) är 0.1.

(a) Kanalens ägare bestämmer sig för att noga följa antalet tittare och likes för en utlagd video med katter. Vad är sannolikheten att de två första tittarna inte ger en like och att de två följande tittarna båda ger en like? (4 p)

(b) Vad är sannolikheten att den 10:e tittaren är den första som ger en like? (4 p)

(c) Vad är sannolikheten att 3 av de 5 första tittarna ger en like? (4 p)

(d) Antag att 120 personer kommer att se en video. Vad är det förväntade antalet likes? Vad är standardavvikelsen för antalet likes? (4 p)

(e) Vad är sannolikheten att fler än 15 personer av de 120 tittarna kommer gilla videon? Du får göra eventuella approximationer om du kan motivera dem. (6 p)

**Uppgift 3** (18 poäng)

Den genomsnittliga mängden bly i blodet för ett lands population är 17 mikrogram/deciliter per person. I ett stickprov från 5 trafikpoliser uppmättes mängden bly enligt tabellen nedan.

| Polis: | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Bly mikrogram/dl: | 30.9 | 27.5 | 26.6 | 17.7 | 35.4 |

(a) Antag modellen $X_1, \ldots, X_5 \sim N(\mu, \sigma)$ för trafikpolisernas mätningar. Skatta $\mu$ med en väntevärdesriktig skattning/estimator. Förklara begreppet väntevärdesriktig estimator. (4 p)

(b) Beräkna ett 95%-igt konfidensintervall för $\mu$ baserat på de 5 mätningarna. Gör en korrekt tolkning av konfidensintervallet som visar att du förstår vad ett konfidensintervall innebär. (7 p)

(c) Testa på 5%-signifikansnivå om trafikpoliser har en *högre* halt bly i blodet jämfört med landets population. Ställ upp hypotester, beräkna teststatistiska och dra slutsats. (7 p)

**Uppgift 4** (16 poäng)

En lärare är intresserad av eventuella skillnader i tentaresultat för studenter med och utan särskilt stöd. Hon tar ett stickprov med 5 studenter utan stöd och 5 studenter med särskilt stöd och kollar upp dessa 10 studenters tentaresultat på en tenta. Hon sammanfattar data i tabellen nedan tillsammans med lite frihetsgrader (df) som hon fått från olika funktioner i R, men hon är osäker på vilka sammanfattningar av data som är relevanta, och hur hon ska undersöka eventuella skillnader mellan grupperna.

| | observationsnummer | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Medelvärde | Stickprovsstandardavvikelse |
| Inget stöd, $x_{1i}$ | 55 | 69 | 54 | 60 | 75 | 62.6 | 9.127 |
| Särskilt stöd, $x_{2i}$ | 44 | 63 | 62 | 49 | 69 | 57.4 | 10.455 |
| Differens, $d_i = x_{1i} - x_{2i}$ | 11 | -6 | -8 | 11 | 6 | 5.2 | 7.791 |
| | | | | | | | |
| $\text{df} = n - 1 = 4$ | | | | | | | |
| $\text{df} = 7.8568$ | | | | | | | |

(a) Hjälp läraren att testa på 5% signifikansnivå om det finns någon skillnad i medelresultat mellan de två studentgrupperna. Ställ upp hypoteser, utför testet och dra slutsatser. Antag att tentaresultaten är normalfördelade. (8 p)

(b) Antag nu att betygsfördelningen för studenter utan stöd är $X \sim N(60, 10)$ (sannolikheten för mer än maxpoängen 100 är extremt liten här, så vi bortser från den). En students mamma lovar hennes dotter att hon ska få 100 kr om hon skriver tentan och sen 5 kronor för varje poäng hon får. Beräkna väntevärde och standardavvikelse för studentens intjänade pengar givet att hon skriver tentan. Vad är sannolikheten att hon tjänar mindre än 300 kr?(8 p)

**Uppgift 5** (24 poäng)

En tidning i en storstad ville undersöka om en restaurangs betyg (`rating` på en skala från 1 till 5) är relaterat till ett högre pris på en standardmiddag (`price`). Man valde slumpmässigt ut 50 restauranger, registrerade deras `rating` och `price` och anpassade sen regressionsmodellen:

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

Den anpassade regressionen ges i utskriften nedan, där vissa tal har tagits bort.

```
Call:
lm(formula = price ~ rating, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-144.342  -32.347    5.586   36.211  161.531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  106.613     23.958
rating        13.746      7.175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.46 on 48 degrees of freedom
Multiple R-squared:  0.07104,   Adjusted R-squared:  0.05168
```

(a) Testa om `rating` är en signifikant förklarande variabel på 10% signifikansnivå. Ställ upp hypoteser, beräkna teststatistikan, ange fördelningen under $H_0$ och utför testet. Dra slutsats. (7 p)

(b) Tidningen får kritik för att de inte tagit hänsyn till restaurangens läge. De samlar därför in ytterligare två förklarande variabler:

- `seaside` som är en binär variabel som är 1 om restaurangen ligger nära havet och 0 annars
- `distance` som mäter avståndet i km från stadskärnan.

Regressionmodellen som anpassas nu är

$$\texttt{price} = \beta_0 + \beta_1 \cdot \texttt{rating} + \beta_2 \cdot \texttt{seaside} + \beta_3 \cdot \texttt{distance} + \varepsilon, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon).$$

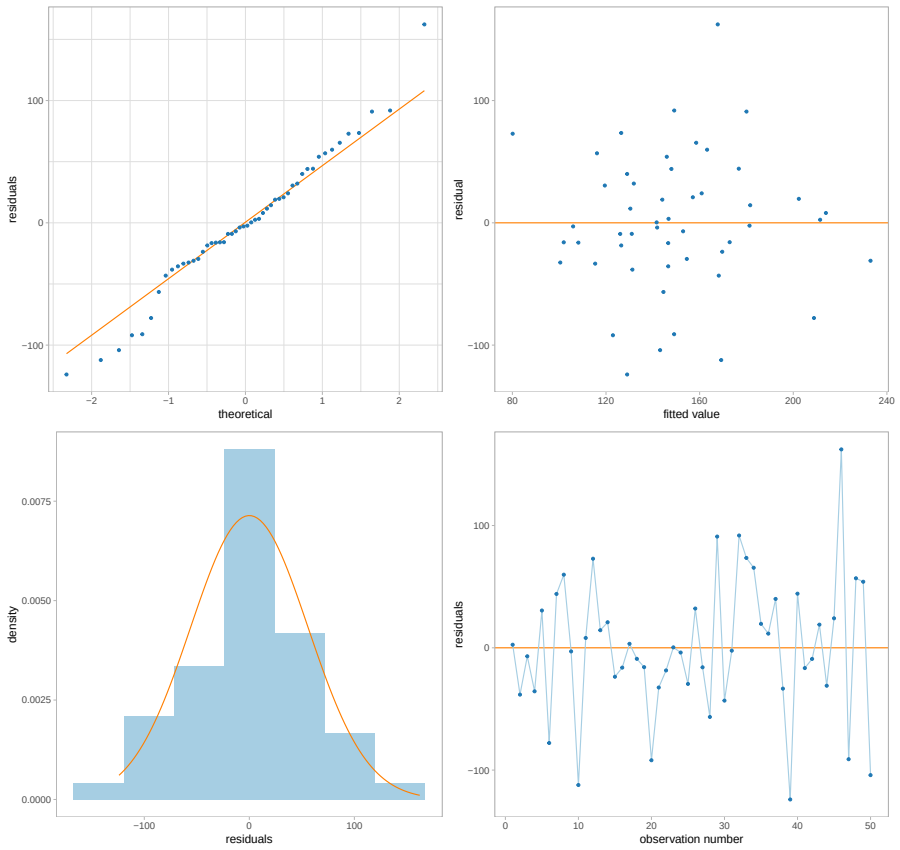
Utskriften nedan ger resultatet från en anpassning av denna multipla regressionsmodell, återigen med vissa saknade värden.
Tolka skattningen av $\beta_3$. Beräkna ett 95%-igt konfidensintervall för $\beta_3$ och använd konfidensintervallet för att testa om $\beta_3$ är en signifikant variabel på 5% signifikansnivå. (9 p)

```
Call:
lm(formula = price ~ rating + seaside + distance, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-124.035  -30.649   -2.643   31.710  162.256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.615     23.179   5.290  3.3e-06 ***
rating        14.804      6.764   2.189  0.03373 *
seaside       93.078              3.142  0.00293 **
distance     -30.700     11.655
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.68 on 46 degrees of freedom
Multiple R-squared:  0.2408,    Adjusted R-squared:  0.1913
```

(c) Utgå från populationsmodellen

$$\texttt{price} = \beta_0 + \beta_1 \cdot \texttt{rating} + \beta_2 \cdot \texttt{seaside} + \beta_3 \cdot \texttt{distance} + \varepsilon, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon).$$

och ange de fyra grundantaganden i populationsmodellen för linjär regression. Beskriv kortfattat för varje antagande om vilken av nedanstående residualplottsfigur som (delvis) kan användas för att undersöka respektive antagande. (8 p)




**Lycka till!**