STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

# Tentamen i Dataanalys och Regression, 4.5 hp

**Kurs**: ST1101, Statistik och Dataanalys I, 15 hp
**Tentamensdatum**: 2024-03-27

Skrivtid: 08.00-12.00 (4 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text. Lexikon. Linjal.
Bifogade hjälpmedel: Formelsamling och statistiska tabeller.

---

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

Svar med fullständiga redovisningar ska lämnas. Svar ges på svenska eller engelska.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga och väl motiverade lösningar.
- 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.

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

Betygsgränser:

| | |
|---|---|
| A: | 90-100 |
| B: | 80-89 |
| C: | 70-79 |
| D: | 60-69 |
| E: | 50-59 |
| Fx: | 40-49 |
| F: | 0-39 |

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

`Lycka till!`

---

**Uppgift 1.** (20 poäng)

(a.) På en skola har de ett dataset där varje observation representerar en elev. Fem av variablerna i detta dataset är följande:

1. Personnummer
2. Namn
3. Kön
4. Ålder
5. Betyg i Engelska på en skala från A till F

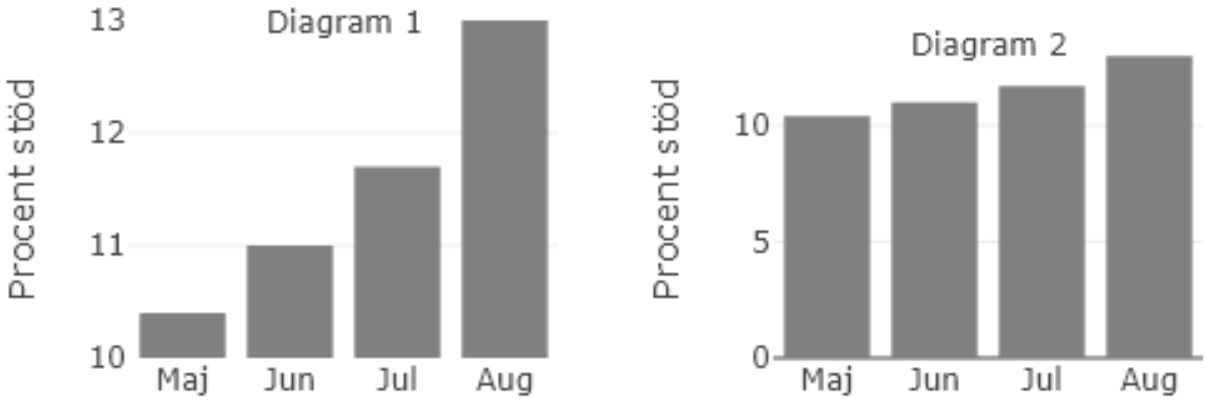
Bland variablerna ovan, välj ut en kategorisk, en numerisk och en ordinal variabel. Motivera ditt svar. (5p)

(b.) Stapeldiagram och histogram är två typer av diagram som ser ungefär likadana ut. Vad är skillnaden mellan stapeldiagram och histogram? (5p)

(c.) Vilken percentil i en fördelning mostsvarar var och en av följande? (5p)

1. Den första kvartilen
2. Den tredje kvartilen
3. Medianen

(d.) En tidning ska publicera en artikel om ett politisk parti som har gått uppåt i opinionen. Tidningen har tillgång till två olika diagram, som båda visar hur partiets opinionssiffror har stigit under de senaste fyra månaderna. Vilket av diagrammen nedan bör tidningen använda om de vill att diagrammet ska leva upp till areaprincipen? Motivera. (5p)




Figur 1: Stöd i opinionen för ett parti

**Uppgift 2.** (20 poäng)

(a.) När du har anpassat (fitted) en regressionsmodell kan du vilja studera residualerna. Vad är en residual, och vilka är två av de modellantaganden som du kan verifiera genom att studera residualerna? (6p)

(b.) En vanlig regressionsmodell anpassas genom minsta kvadrat-metoden. På vilket sätt är minsta kvadrat-metoden kopplad till residualerna? (4p)

(c.) Du ska göra en regressionsmodell som predikterar inkomst med hjälp av en eller flera förklaringsvariabler. Du överväger två alternativa modeller.

1. Modell 1 inkluderar enbart förklaringsvariabeln ålder.
2. Modell 2 inkluderar både förklaringsvariabeln ålder och förklaringsvariabeln utbildning, som anger hur många år en person har studerat.

Nämn två metoder som du kan använda för att jämföra de båda modellerna och avgöra vilken som är lämpligast. (6p)

(d.) En analytiker vill ta reda på medelinkomsten i Sverige, och har i det syftet samlat in inkomstuppgifter från 1000 slumpvis valda invånare i Sollentuna kommun. Analytikerns chef säger att undersökningen måste göras om eftersom urvalsmetod som användes kommer att medföra bias. Vad är bias, och har chefen rätt? Motivera svaret. (4p)

**Uppgift 3.** (20 poäng)

Ett e-handelsbolag som säljer skor på nätet har haft en reklamkampanj på sociala medier. I reklamkampanjen har de visat 3 olika annonser som vi kallar annons A, annons B och annons C.

Resultatet av varje annonsvisning var antingen att den ledde till ett köp eller att den *inte* ledde till ett köp. Resultatet av alla annonsvisningar visas i korstabellen nedan.

| | | Köp | |
|---|---|---|---|
| | | Ja | Nej |
| Annons | Annons A | 506 | 31500 |
| | Annons B | 532 | 33300 |
| | Annons C | 287 | 28800 |

Variabeln *Annons* anger vilken av annonserna som visades vid ett visst tillfälle. Variabeln *Köp* anger om annonsvisningen resulterade i ett köp eller inte.

(a.) Hur många annonsvisningar gjordes totalt under kampanjen? (2p)

(b.) Räkna ut marginalfördelningarna (marginal distributions) för båda variablerna. Uttryck marginalfördelningarna i procent. Tolka marginalfördelningarna. (4p)

(c.) Företaget vill i efterhand utvärdera hur väl de olika annonserna fungerade. För att göra detta, bör vi räkna ut fördelningen av variabeln *Köp* betingad på variabeln *Annons*, eller bör vi räkna ut fördelningen av variabeln *Annons* betingad på variabeln *Köp*? Motivera ditt svar. (4p)

(d.) Räkna ut den betingade fördelning som du föreslog i deluppgift **c**. Uttryck svaren i procent. (8p)

(e.) Tolka resultatet från deluppgift **d**. (2p)

**Uppgift 4.** (20 poäng)

En lärare mäter sina elevers läshastighet. Varje elev läser en text som innehåller ett visst antal ord, och läraren mäter antalet sekunder som det tar elever att läsa texten. Med hjälp av resultaten anpassar läraren regressionsmodellen

$$\hat{y} = 3.2 + 0.34x,$$

där $x$ är antalet ord i texten som en elev läser och $\hat{y}$ är det estimerade tiden i sekunder som det tar att läsa texten.

(a.) Tolka modellens koefficienter (coefficients), dvs interceptet och lutningskoefficienten (slope coefficient). (4p)

(b.) Vi har följande information: $r_{xy} = 0.9, s_x = 90, s_y = 34, \bar{x} = 720, \bar{y} = 248$.
Förklara notationen, dvs förklara vad $r_{xy}$, $s_x$, $s_y$, $\bar{x}$ och $\bar{y}$ står för.
Verifiera sedan med hjälp av denna information att $b_0 = 3.2$ och att $b_1 = 0.34$. (4p)

(c.) Vissa av eleverna har gått en kurs i snabbläsning. Läraren lägger till en dummyvariabel som har värdet 1 för elever som har deltagit i snabbläsningskursen och 0 för elever som inte har gått kursen. Modellen blir nu

$$\hat{y} = 4.1 + 0.4x_1 - 24x_2,$$

där $x_1$ är antalet ord i texten, $x_2$ är dummyvariabeln för deltagande i snabbläsningskursen, och $\hat{y}$ är den estimerade tiden i sekunder som det tar att läsa texten.
Tolka modellens samtliga tre koefficienter. Kom ihåg att vi nu har en multipel linjär regressionsmodell. (4p)

(d.) I en tidigare studie om läshastighet användes regressionsmodellen

$$\widehat{log(y)} = 0.5 + 0.007x,$$

där $x$ är antalet ord i texten och $\widehat{log(y)}$ är estimatet av den logaritmerade tiden i sekunder som det tar att läsa texten. Med logaritmen avser vi den naturliga logaritmen, dvs logaritmen med bas e. Hur många sekunder estimerar denna modell att det tar att läsa en text som består av 700 ord? (4p)

(e.) Vi har följande information om en regressionsmodell:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 30000$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 25000$$

Räkna ut R-squared för modellen. (4p)

**Uppgift 5.** (20 poäng)

Ett stort antal löpare deltog i ett maratonlopp. Tiderna som det tog för deltagarna att springa loppet följde en normalfördelning med medelvärdet 235 minuter och standardavvikelsen 40 minuter.

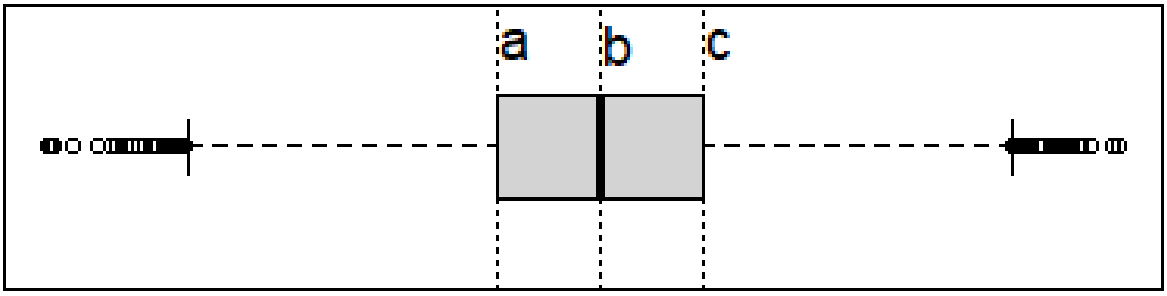
(a.) Leah sprang loppet på 255 minuter. Hur stor andel av löparna hade en *lägre* tid än Leah? (5p)

(b.) Hur många minuter tog loppet att springa för en löpare vars tid befann sig vid den 10:e percentilen i fördelningen? (5p)

(c.) Beräkna kvartilavståndet (IQR) för fördelningen av tider i loppet. (5p)

(d.) Figuren nedan visar fördelningen av löparnas tider i minuter i form av ett låddiagram (boxplot). Vilka tider, räknat i minuter, representerar a, b och c?

(Om du inte vet vilka tider, räknat i minuter, som representeras av a, b och c, kan du då förklara i ord vad a, b och c i låddiagrammet representerar?) (5p)




Figur 2: Fördelning av löpartider i minuter