STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2023-09-22

Skrivtid: 08.00-12.00 (4 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text. Lexikon.
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. Lösningsförslag läggs ut på Athena efter tentamen i samband med rättningen.

`Lycka till!`

Uppgift 1. (20 poäng)

(a.) Nämn en skillnad mellan kategoriska variabler och numeriska variabler. Ge ett exempel på en kategorisk respektive en numerisk variabel. (5p)

(b.) Vad är en outlier? (5p)

(c.) Vad är bias? Ge ett exempel, verkligt eller påhittat, på hur bias kan uppkomma när data samlas in. (5p)

(d.) Vad är en dold variabel (lurking variable)? Ge ett exempel på ett scenario, verkligt eller påhittat, där en dold variabel har betydelse. (5p)

Uppgift 2. (20 poäng)

(a.) Vad är en förklaringsvariabel (explanatory variable) respektive en responsvariabel (response variable) i en regressionsmodell? (4p)

(b.) Vad är en residual i en regressionsmodell? Hur beräknas den? (4p)

(c.) När du har anpassat (fitted) en regressionsmodell kan du studera residualerna för att se om de modellantaganden (model assumptions) som gäller för en regressionsmodell är uppfyllda. Nämn två sådana antaganden. (6p)

(d.) En grupp ingenjörer diskuterar vilken av två regressionsmodeller som är bäst. En ingenjör föredrar modell 1, som har lägre RMSE när den utvärderas genom korsvalidering (cross validation). En annan ingenjör invänder och hävdar att en modell med högre R-kvadrat (R-squared) alltid är bättre, och eftersom den mer flexibla modell 2 har högre R-kvadrat bör den väljas. Har den andra ingenjören rätt i sitt resonemang on R-kvadrat? Motivera ditt svar. (6p)

Uppgift 3. (20 poäng)

En butikskedja har tre butiker, som ligger i Stockholm, Göteborg och Malmö. Butikskedjan har gjort en kundundersökning, där kunder i de tre butikerna har tillfrågats om de är nöjda eller inte med sitt senaste köp. Resultatet av undersökningen visas i tabellen.

		Nöjd	
		Ja	Nej
Stad	Stockholm	138	60
	Göteborg	101	47
	Malmö	60	39

Variabeln *Stad* anger i vilken butik en kund har handlat och variabeln *Nöjd* anger om kunden är nöjd eller inte.

(a.) Hur många kunder totalt svarade på frågan? (2p)

(b.) Hur många kunder i undersökningen hade handlat i Malmö-butiken och var missnöjda med sitt senaste köp? (2p)

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

(d.) Räkna ut fördelningen av variabeln *Nöjd* betingad (conditioned) på variabeln *Stad*. Ange de betingade fördelningarna i procent. (8p)

(e.) Tolka resultatet från deluppgift **d** i termer av skillnader i kundnöjdhet. (4p)

Uppgift 4. (20 poäng)

Ett antagningsprov till en utbildning i journalistik mäter ordförståelse. Baserat på ett stort antal provresultat vet vi att provresultaten är normalfördelade. Medelvärdet är 55 poäng, och standardavvikelsen är 8 poäng.

(a.) Anna gör provet och får 59 poäng. Hur stor andel av alla som har gjort provet har fått *högre* poäng än Anna? (4p)

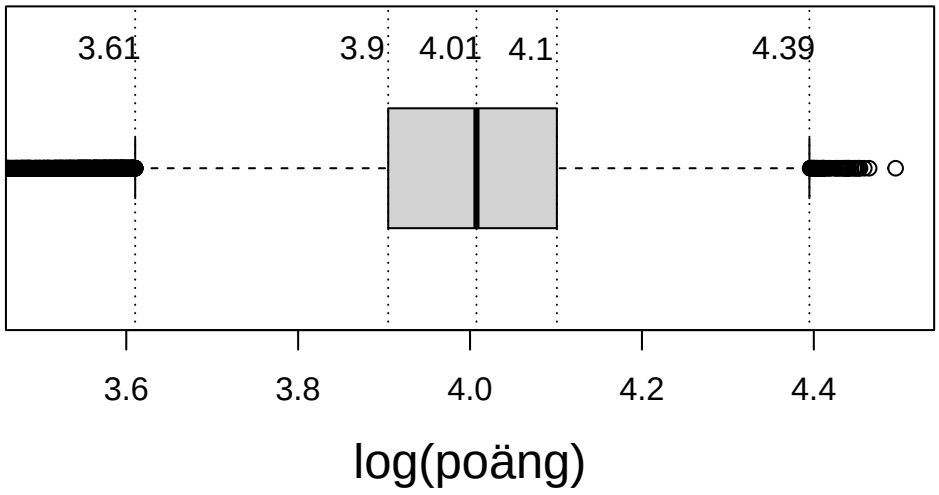
(b.) För att bli antagen till utbildningen krävs att ditt provresultat tillhör de högsta 10 procenten, dvs att resultatet ligger vid den 90:e percentilen eller högre. Hur många poäng behöver du för att uppnå detta? (Avrunda svaret uppåt till närmaste heltal.) (4p)

(c.) Räkna ut kvartilavståndet (IQR) för fördelningen av provresultat. Gör detta med hjälp av normalfördelningen. (5p)

(d.) Låddiagrammet (box plot) är ett sätt att illustrera fördelningen av provresultaten. Skalan i diagrammet är *log(poäng)*, alltså logaritmerade poäng.

Räkna återigen ut kvartilavståndet (IQR) för poäng-fördelningen, denna gång med hjälp av informationen som ges i diagrammet. Kvartilavståndet ska anges i original-skala, så kvartilerna måste transformeras från den logaritmerade skalan. Notera att värden för de streckade vertikala linjerna är inskrivna i diagrammet. (7p)

Notera: Eftersom du räknade ut kvartilavståndet även i deluppgift **c** bör du här få ungefär samma svar, men eftersom siffrorna i diagrammet är avrundade blir svaret inte exakt samma.



Uppgift 5. (20 poäng)

En glasskiosk har samlat data som visar medeltemperaturen i grader Celsius och antalet sålda glassar för ett antal dagar. De har använt denna data till att anpassa regressionsmodellen

$$\hat{y} = 132 + 17x,$$

där x är medeltemperaturen i grader Celsius en viss dag och $\hat{y}$ är det estimerade antalet sålda glassar den dagen.

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

(b.) Hur många glassar estimerar modellen att glasskiosken säljer en dag då det är 24 grader Celsius? (4p)

(c.) Vi har följande information: $r_{xy} = 0.7, s_x = 7.8, s_y = 190, \bar{x} = 18, \bar{y} = 439$.

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 $b_0 = 132$ och att $b_1 = 17$, avrundat till heltal. (4p)

(d.) Vi lägger till dummy-variabeln *regn*, som har värdet 1 om det regnade en viss dag och annars 0. Modellen blir nu.

$$\hat{y} = 290 + 8x_1 - 206x_2,$$

där x_1 är temperaturen i grader Celsius och x_2 är dummyvariabeln för regn.

Tolka modellens samtliga tre koefficienter, dvs interceptet och de två lutningskoefficienterna. (4p)

(e.) Om det regnar en dag har x_2 värdet 1 och annars värdet 0. Formulera regressionsmodellen så som den skulle se ut om kodningen var den omvända, dvs om dagar med regn kodades som 0 och regnfria dagar som 1. Visa med ett exempel att prediktionerna blir samma oavsett hur dummyvariabeln kodas. (4p)