STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2023-11-02

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 två typer av diagram som kan användas för att illustrera fördelningen för kategoriska variabler. Nämn även två typer av diagram som kan användas för att illustrera fördelningen för numeriska variabler. (5p)

(b.) Rita en unimodal, en bimodal, och en uniform fördelning. Skriv ut vilken graf som illustrerar vilken fördelning. Det behöver inte vara snyggt ritat, men för poäng måste det synas tydligt vilken typ av fördelning varje graf illustrerar. (5p)

(c.) Tre centralmått för en fördelning är typvärde (mode), median och medelvärde (mean). Om en fördelning är skev åt höger (right skewed), vilket av dessa centralmått kan förvantas ha högst värde? Motivera ditt svar, och rita gärna en bild. (5p)

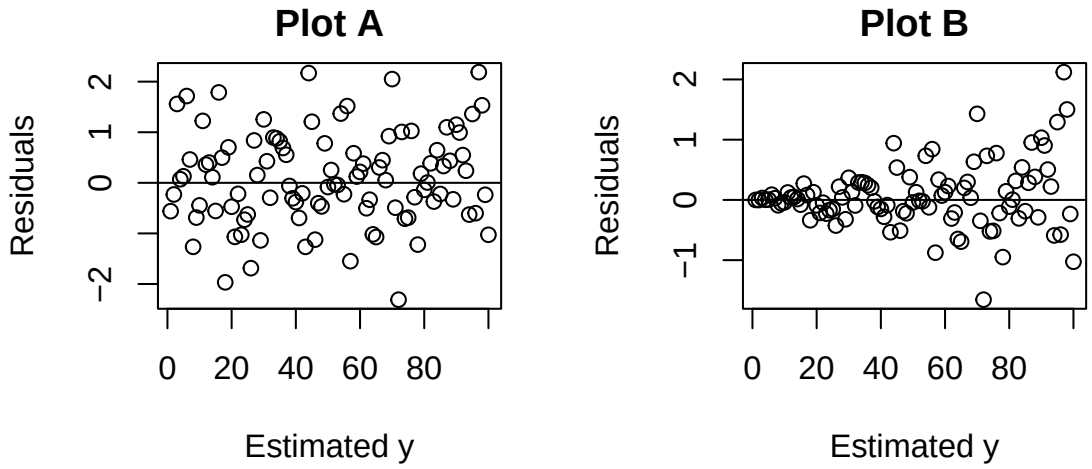
(d.) Nämn en fördel med studier i form av experiment jämfört med observationsstudier. (5p)

Uppgift 2. (20 poäng)

(a.) När vi anpassar en enkel linjär regressionsmodell, det vill säga när vi bestämmer vilka värden b_0 och b_1 ska ha, gör vi det med hjälp av Minsta Kvadratmetoden (Least Squares Method). Vad menas med att vi använder Minsta Kvadratmetoden? (5p)

(b.) Vad menar vi när vi säger att sambandet mellan responsvariabeln och förklaringsvariabeln i en enkel linjär regressionsmodell inte är statistiskt signifikant? (5p)

(c.) Nedan ser vi residualgrafer för två olika linjära regressionsmodeller. Vilken av graferna kommer från en modell där vi ser att alla våra modellantaganden inte är uppfyllda? Vilket modellantagande är inte uppfyllt? (5p)



(d.) En regressionsmodell bör varken ha för hög eller för låg flexibilitet. Förklara varför. (5p)

Uppgift 3. (20 poäng)

Ett nytt läkemedel prövas ut med hjälp av ett antal patienter. En grupp patienter i undersökningen fick läkemedlet medan en annan grupp patienter fick ett placebopiller, dvs ett piller som är verkningslöst. Patienterna visste inte själva om de fick det riktiga läkemedlet eller placebopillret.

Efter en tids behandling meddelade varje patient om de mådde bättre eller inte. Resultatet visas i tabellen nedan.

		Effekt	
		Ja	Nej
Grupp	Läkemedel	178	70
	Placebo	121	57

Variabeln *Grupp* anger om en patient fick läkemedlet eller placebopillret, och variabeln *Effekt* anger om patienten mådde bättre eller inte.

(a.) Vad kallas denna typ av tabell, och vilken typ av variabler är det vi studerar med den här sortens tabell?(4p)

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

(c.) Räkna ut fördelningen av variabeln *Effekt* betingad (conditioned) på variabeln *Grupp*. Ange de betingade fördelningarna i procent. (8p)

(d.) Tolka resultatet från deluppgift **c**. Vad säger resultatet om läkemedlets effekt? (4p)

Uppgift 4. (20 poäng)

En butiksägare väger de äpplen som kommit med den senaste leveransen. Det visar sig att äpplenas vikt följer en normalfördelning med medelvärdet 200 gram och standardavvikelsen 40 gram.

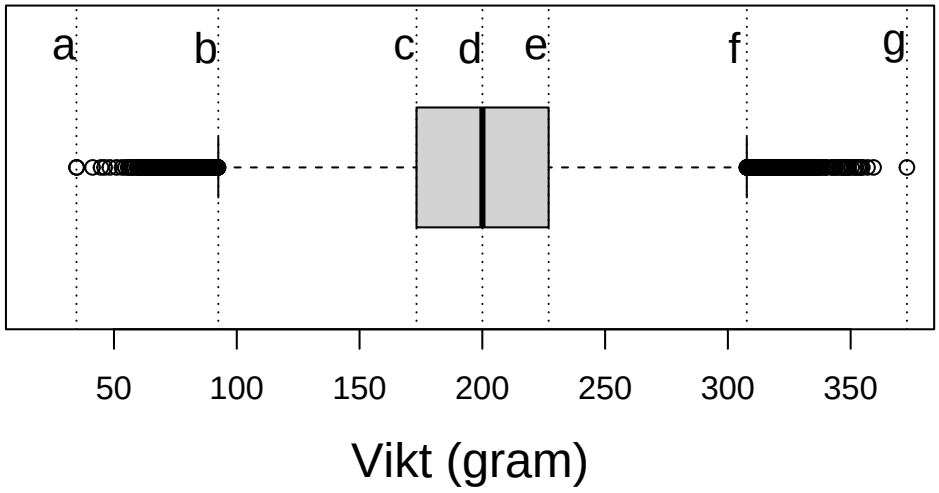
(a.) Leverantören hävdar att nästan alla äpplen i leveransen väger mer än 140 gram. För att undersöka hur väl detta stämmer, använd normalfördelningen och räkna ut hur stor andel av alla äpplen i leveransen som väger mer än 140 gram. (5p)

(b.) Handlaren sorterar ut de 10 procent äpplen som väger mest, för att sälja dessa dyrare. Hur mycket måste ett äpple väga för att tillhöra dessa 10 procent, dvs vad måste ett äpple väga som minst för att vikten ska ligga över den 90:e percentilen? (5p)

(c.) Hur stor andel av alla äpplen har en vikt som avviker med högst 40 gram från medelvärdet? Hur stor andel har en vikt som avviker med högst 80 gram från medelvärdet? (5p)

(d.) Låddiagrammet (box plot) är ett sätt att illustrera viktfördelningen för de äpplen som levererades.

Vad betecknar var och en av bokstäverna a, b, c, d, e, f och g i diagrammet? (Du behöver inte läsa av siffrorna eller räkna ut något i denna deluppgift) (5p)



Uppgift 5. (20 poäng)

En badvakt har samlat in data över antalet personer som badar på en badstrand dag för dag. Han har använt denna data till att anpassa regressionsmodellen

$$\hat{y} = 29 + 3.2x,$$

där x är vattentemperaturen i grader Celsius en viss dag och $\hat{y}$ är det estimerade antalet personer som badar.

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

(b.) En viss dag estimerade badvaktens modell att 95 personer skulle bada på stranden. Ungefär vilken temperatur var det den dagen? (4p)

(c.) Vi har följande information: $r_{xy} = 0.4, s_x^2 = 37.21, s_y^2 = 2401, \bar{x} = 19, \bar{y} = 90$.

Förklara vad r_{xy}, s_x^2 och s_y^2 står för.

Verifiera sedan med hjälp av den givna information att $b_0 \approx 29$ och att $b_1 \approx 3.2$. (4p)

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

$$\hat{y} = 72 + 1.2x_1 - 26x_2,$$

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

Tolka modellens samtliga tre koefficienter b_0, b_1 och b_2. (4p)

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

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

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 9000$$

Hur stor del av responvariabelns variation förklarar regressionsmodellen? (4p)