STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
TENTAMEN
21 febuari 2024

Tentamen i Statistisk analys

21 febuari 2024 kl. 8–13

Examinator: Tom Britton, tel. 08-16 45 34, tom.britton@math.su.se
Tillåtna hjälpmedel: Formel- och tabellsamling (delas ut) och miniräknare.
Återlämning: Tentan kommer vara rättad senast två veckor efter tentamensdagen. När tentan är rättad meddelas detta via kursforumet och därefter kan tentan hämtas ut vid matematikexpeditionen.

Varje korrekt löst uppgift ger 10 poäng. Gränsen för godkänt är preliminärt 30 poäng. För D krävs 34 p, för C 40 p, för B 48 p och för A krävs 54 p (alla gränser gäller inkl ev bonuspoäng). Texten ska vara väl läsbar och resonemang ska vara klara och tydliga.

Det skall tydligt framgå hur beräkningar gjorts. Kommunikation med andra personer är **ej** tillåtet och kommer anmälas vid uppdagande.

Uppgift 1

Nedan följer 5 påståenden att svara sant eller falskt på (eller ingenting om man inte vet). Korrekt svar på respektive påstående ger 2p, fel svar ger -2p och inget svar ger 0p (om totalsumman skulle bli negativ sätts poängen till 0).

a) Parametriska metoder som använder normalfördelningen kan vara acceptabla även om ursprungliga data inte är normalfördelade, t ex om fördelningen inte är tjocksvansad och när antalet observationer är tillräckligt många.

b) I Bayesiansk statistik utgår man från en statistisk modell för data givet parametern (likelihooden), samt en apriorifördelning för sin kunskap (eller okunskap) om parametern. Efter att data samlats in beräknar man aposteriorifördelningen, som är proportionell mot produkten av apriorifördelningen och likelihooden, som blir den uppdaterade kunskapen om parametern.

c) När man genomför ett hypotestest väljs en lämplig signifikansnivå α, t ex 1% eller 5%. Detta värde anger vald sannolikhet att förkasta nollhypotesen H_0 när H_0 i själva verket är sann.

d) De två viktigaste egenskaperna för en punktskattning θ^* av parametern θ är 1) väntevärdesriktighet (dvs $E(\theta^*) = \theta$), och 2) att $var(\theta^*)$ är liten.

e) Median är det mest använda lägesmåttet.

Uppgift 2

Kontaktstudier mäter hur många kontakter (definierad som samtal inom armlängds avstånd i minst 2 minuter) slumpvis utvalda individer har under en dag. Tidigare studier har visat att individer tenderar att ha färre kontakter ju äldre man blir.

I en ny studie beräknades genomsnittligt antal kontakter bland 50 individer för $n = 6$ olika åldrar. Resultatet blev som följer (x=ålder, y=genomsnittligt antal kontakter): (20, 16.3), (30, 16.0), (40, 14.3), (50, 12.9), (60, 10.8), (70, 6.0).

a) Rita upp data i ett diagram. Ansätt linjär regression och definiera modellen, samt kommentera modellen i relation till diagrammet. (3 p)

b) Skatta modellens 3 (!) parameterar. Du får använda dig av följande beräknade summor: $\sum_i x_i = 270,\ \sum_i y_i = 76.3,\ \sum_i x_i^2 = 13900,\ \sum_i y_i^2 = 1045.23,\ \sum_i x_i y_i = 3091.$ (3 p)

c) Testa frågeställningen som nämns ovan, dvs om det finns någon tendens att antal kontakter ändras med åldern eller inte. Gör ett tvåsidigt test på 95%-nivån. (4 p)

Uppgift 3

I januari 2021 införde rektor Astrid Söderberg-Widding vid Stockholms universitet projektet *Ekonomi i balans* i syfte att institutionerna vid SU skulle strama upp sin ekonomi. Här kommer det ekonomiska utfallet (i enhet MSek) för 8 institutioner för åren 2020 (dvs före projektets start) och 2021 (dvs efter projektets start): (-3.6, -2.0), (1.8, 2.5), (-5.4, -3.2), (3.1, 2.4), (-0.8, -0.5), (-4.1, -2.8), (-0.4, 1.1), (0.3, 0.5).

Det går bra att anta att *förändringen* av resultat mellan åren är oberoende och likafördelade (med snäll fördelning) mellan institutionerna. (Obs projektet *Ekonomi i balans* har existerat men data är fingerat.)

a) Testa om projektet hade positiv effekt det första året, dvs om institutionerna fick signifikant bättre resultat 2021 jämfört med året innan. Du får välja mellan ensidigt eller tvåsidigt test, och lämplig signifikansnivå, men dessa ska redovisas. (5 p)

b) Konstruera att 95% konfidensintervall för effekten av projektet. (4 p)

c) Finns det något i givna antaganden som du tycker kan ifrågasättas? (1 p)

Uppgift 4

I en kontaktstudie ombads deltagande individer (i en viss åldersgrupp) ange hur många kontakter (definierat som i Uppgift 2) de haft under två separata dagar. En viktig frågeställning inom sociologi är att utreda om variationen i antal kontakter som individer har huvudsakligen är rent slumpmässig, eller om vissa individer tenderar att ha konsekvent fler och andra konsekvent färre kontakter olika dagar. Följande antal kontakter erhölls från $n = 10$ individer (dag 1, dag 2): (12, 18), (4,9), (6, 8), (14, 10), (13, 29), (7,5), (10, 9), (8,13), (12, 14) och (5,6).

Till er hjälp kan ni använda er av följande summor: $\sum_i x_i = 91,\ \sum_i y_i = 121,\ \sum_i x_i^2 = 943,\ \sum_i y_i^2 = 1917,\ \sum_i x_i y_i = 1244.$

a) Illustrera datamaterialet i ett spridningsdiagram. (2 p)

b) Beräkna Pearson-korrelationen r_{xy}. (3 p)

c) Genomför ett test av om olika individer verkar ha lika många kontakter i genomsnitt, eller om det verkar finnas en samband mellan en individs kontakter de olika dagarna (dvs en individeffekt). Använd 5% signifikansnivå. (5 p)

Uppgift 5

Samma datamaterial och allmänna frågeställning som i Uppgift 4.

a) Ange ett argument till varför det kanske inte är lämpligt att använda parametrisk inferens (Perasonkorrelationen) för detta datamaterial. (2 p)

b) Beräkna därför Spearmankorrelationen och genomför ett test för om det finns en individeffekt eller inte, dvs om en individs kontakter de olika dagarna verkar vara oberoende eller beroende. Använd 5% signifikansnivå. (8 p)

Uppgift 6

Antag att flygpassagerare (inklusive handbagage och kläder) tidigare hade en medelvikt på $\mu_0 = 80$ kg med en standardavvikelse på $\sigma = 10$ kg (inkluderar även könsskillnader). Man vill nu undersöka om denna vikt gått upp eller inte (ensidigt test) och för detta ska man väga n individer. Du kan fortfarande anta att $\sigma = 10$ kg, men frågan är alltså om $\mu = 80$ fortfarande, eller om $\mu > 80$.

a) Beskriv hur man ska testa nollhypotesen från data $x_1, \ldots, x_n$ på 5%-nivån. (3 p)

b) Antag att man tror att medelvikten ökat med 2kg (och således att $\mu = 82$). Hur stort måste n väljas för att kunna upptäcka denna ökning med 80% sannolikhet? (En s.k. power-beräkning) (7 p)

Lycka till!