STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
TENTAMEN
23 augusti 2022

Tentamen i Statistisk analys

23 augusti 2022 kl. 14–19

Examinator: Tom Britton, tel. 08-16 45 34, tom.britton@math.su.se
Tillåtna hjälpmedel: Formel- och tabellsamling och miniräknare.
Återlämning: Tentan kommer vara rättad senast 30/8 2022 och återfinns då vid matematikexpeditionen (Obs! Notera att institutionen flyttat till Albano!).

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. Texten ska vara väl läsbar och resonemang ska vara klara och tydliga.

Lösningar till uppgifterna måste göras självständigt och 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) Skattningar (vid opinionsmätningar) av om ett politiskt part gått framåt eller ej har större osäkerhet om den eventuella uppgången bedöms mot föregående opinionsmätning jämfört med om den bedöms mot senaste valresultatet (som kan betraktas som en totalundersökning).

b) Ett stort endimensionellt datamaterial med många olika möjliga observationsvärden illustreras väl med ett histogram.

c) Ett givet hypotestest (dvs för en given statistisk situation och given stickprovstorlek) för nollhypotesen $H_0 : \alpha = 10$ har större chans att förkastas om det sanna parametervärdet är $\alpha = 11$ jämfört med om det sanna parametervärdet är $\alpha = 12$.

d) Antalet frihetsgrader vid statistiska test är ofta antalet observationer subtraherat med antal parametrar som skattas.

e) Regressionsanalys baseras på att de olika observationerna är oberoende men inte nödvändigtvis likafördelade.

Uppgift 2

Ett stickprov av storlek 15 samlades in från en viss ”snäll” fördelning med väntevärde μ. Utfallet blev sådant att $\bar{x} = (1/15)\sum_{i=1}^{15} x_i = 6.71$ och $s^2 = (1/14)\sum_{i=1}^{15}(x_i - \bar{x})^2 = 0.074$

a) Skapa ett 2-sidigt 95%-konfidensintervall för μ. (4 p)

b) Testa den ensidiga hypotesen $H_0 : \mu = 7$ mot alternativet $H_1 : \mu < 7$ på 99%-nivån. (4 p)

c) Ge en kort motivering till varför t-fördelning får användas även om fördelningen för X ej är normalfördelad.

Uppgift 3

För att bilda sig en uppfattning om hur årsmedeltemperaturen sjunker norrut i Sverige används enkel linjär regression. Årsmedeltemperaturen har samlats in från $n = 11$ orter i Sverige med känd lattitud: Jokkmokk: latt: 66.6, medeltemp: -0.6, Umeå: 63.5, 4.0, Östersund: 63.1, 4.2, Gävle: 60.4, 5.8, Karlstad: 59.2, 7.0, Stockholm: 59.3, 7.6, Jönköping: 57.4, 6.0, Visby: 57.6, 7.6, Göteborg, 57.8, 7.7, Kalmar: 56.7, 7.5, Lund: 55.7, 8.5. Följande sammanfattande storheter beräknades från data (x är lattitud och y årsmedeltemperatur): $\sum_i x_i = 657.3$, $\sum_i x_i^2 = 39389.45$, $\sum_i y_i = 65.3$, $\sum_i y_i^2 = 455.95$ samt $\sum_i x_i y_i = 3820.38$.

a) Skatta och ange ett 99% konfidensintervall för hur mycket medeltemperaturen sjunker per lattitud (inom Sverige). Obs: endast konfidensintervall för efterfrågad storhet ska anges. (En korrekt lösning under antagandet att $\sigma = 1$ ger 3p.) (5 p)

b) Vilken förklaringsgrad R^2 har modellen? (2 p)

c) Skatta medeltemperaturen för Bollnäs med lattitud 61.34 (endast punktskattning). (3 p)

Uppgift 4

50 heltidsarbetande personer vardera från Stockholm, Göteborg och Malmö tillfrågades om färdmedel till jobbet (påhittad data). Alternativen som gavs var: gå, cykel, bil eller kollektivtrafik (och ingen hade annat färdmedel). Utfallet blev som följer:

Stad	Gå	Cykel	Bil	Kollektivt	Summa
Stockholm	6	12	13	19	50
Göteborg	10	8	20	12	50
Malmö	8	19	15	8	50
Antal	24	39	48	39	150

a) Undersök om populationerna i de tre städerna kan antas välja färdmedel på liknande sätt, eller om typ av färdmedel skiljer sig signifikant åt. Testa på 5%-nivån. (7 p)

b) Antag i stället att du endast var intresserad av de två första städerna (Stockholm och Göteborg) och endast huruvida bil används eller ej (slå således ihop övriga svarsalternativ). Ange en skattning och 95% konfidensintervall för skillnaden i användandet av bil mellan Stockholm och Göteborg. (3 p)

Uppgift 5

Vid en undersökning av sexualvanor vid en ungdomsklinik angav de svarande hur många partners de hade haft sex med det senaste året. Man erhöll svar från 8 män och 10 kvinnor. Antal partners som de 8 männen angav var: 2, 6, 9, 0, 2, 0, 13, 0. För de 10 kvinnorna blev svaren 1, 3, 1, 0, 4, 1, 3, 3, 1, 7 (påhittade data). En av frågorna som man vill besvara är om män och kvinnor verkar ha samma *fördelning* för antal sexpartners.

a) Vilka skäl finns att använda icke-parametriska metoder i detta fall? Motivera. (2 p)

b) Hur ska du hantera att det finns observationer med samma numeriska värden? (2 p)

c) Använd lämplig icke-parametrisk metod för att på 95%-nivån testa hypotesen att män och kvinnor (för populationsgruppen som besöker ungdomskliniker!) har samma fördelning av antal partners. (6 p)

Uppgift 6

En okänd andel p i en befolkningsgrupp besitter ett visst genetiskt anlag. Ett stickprov av 50 st slumpvis utvalda individer som inte är släkt med varandra undersöks om de innehar anlaget eller ej, vilket således kommer resultera i att X individer besitter anlaget. Målet med denna uppgift består i att ansätta en Bayesiansk analys av p baserat på det insamlade datamaterialet.

a) Vad är ett rimligt antagande om apriorifördelning för p? Motivera ditt val. Ange även vilken fördelning X har givet värdet på p. (3 p)

b) Antag att $X = 17$ observeras (dvs 17 av de 50 undersökta individerna visar sig besitta anlaget). Härled ett uttryck som aposteriorifördelningen för p är proportionell mot. Notera att du inte behöver härleda konstanter som gör att aposteriorifördelningen integrerar sig till 1 - det räckar att ange de för fördelningen relevanta storheterna. (4 p)

c) En klass av fördelningar för en slumpvariabel på $[0, 1]$-intervallet kallas betafördelning. Mer precist säger man att en sådan variabel Z är betafördelad, $Z \sim Beta(m, n)$, när den har tätheten $f_Z(z) = \beta(m, n)z^{m-1}(1 - z)^{n-1}$ för $0 \leq z \leq 1$ ($f_Z(z) = 0$ utanför detta intervall), där $\beta(m, n)$ är den s.k. betafunktionen som inte är särskilt viktig, men som gör att tätheten integrerar sig till 1. Denna slumpvariabel har f.ö. väntevärde $m/(m+n)$. (T ex gäller att $m = n = 1$ ger likformig fördelning.) Använd dessa resultat för att ange vilken aposteriorifördelning p har, samt aposteriofördelnings väntevärde. Hur relaterar detta till vad de flesta skulle ha trott om p (dvs ML-skattningen)? (3 p)

Lycka till!