STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
TENTAMEN
22 augusti 2023

Tentamen i Statistisk analys

22 augusti 2023 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 fredag 25/8 2023 och återfinns därefter 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) Ett 95% konfidensintervall är bredare än motsvarande 99% konfidensintervall.

b) Arean av en stapel i ett histogram är proportionell mot andelen observationer som ligger i motsvarande intervall.

c) En referensvariabel r för en parameter θ ska ha ett numeriskt värde som således inte innehåller θ.

d) I regressionsanalys får fördelningen för Y_i bero på x_i men däremot inte på övriga Y_j observationer eller x_j-värden.

e) Icke-parametriska metoder kan användas även om data *är* normalfördelat men ger då lägre styrka än parameteriska metoder.

Uppgift 2

Vikten på 10 slumpvis valda trädgårdsjordsäckar av ett visst fabrikat uppmättes: 19.7, 20.3, 20.1, 20.5, 19.9, 20.0, 20.3, 20.6, 20.1, 20.2. Summan av vikterna och deras kvadratsumma ges av: $\sum_{i=1}^{10} x_i = 201.7$ och $\sum_{i=1}^{10} x_i^2 = 4068.95$.

a) Gör ett 90% konfidensintervall för genomsnittliga vikten μ på detta fabrikats jordsäckar. (6 p)

b) Testa på 95%-nivån om μ överstiger 20 kg eller inte. (4 p)

Uppgift 3

Man vill undersöka hur årsmedeltemperaturen sjunker norrut i Sverige och använder för detta 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) Vad är förklaringsgraden R^2 för modellen? (2 p)

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

Uppgift 4

En undersökning vill studera och jämföra antalet fågelungar som två sorters änder får. 10 kullar observerades av respektive art med följande utfall.

Art 1: 4, 6, 3, 6, 5, 4, 2, 3, 3, 4.

Art 2: 2, 5, 3, 2, 4, 3, 2, 2, 4, 3.

Man kan anta att antalet ungar man får är Poissonfördelat med respektive parametrar (=väntevärde) λ_1 and λ_2. Som hjälp kan ni använda följande summor: $\sum_{i=1}^{10} x_i = 40$ och $\sum_{i=1}^{10} x_i^2 = 176$ för art 1 samt $\sum_{i=1}^{10} y_i = 30$ och $\sum_{i=1}^{10} y_i^2 = 100$ för art 2.

a) Härled ML-skattningen för λ_1 och λ_2 på lämpligt sätt och ange respektive numeriska skattning. (4 p)

b) Testa (på lämplig nivå) hypotesen att de två arterna har samma väntevärde (för detta behöver inte deluppgift A användas utan här får ni använda approximationen att medelvärden av flera oberoende och likafördelade Poissonvariabler är normalfördelade). (6 p)

Uppgift 5

En sexualvaneundersökning vid en ungdomsklinik ber de svarande (bl a) att ange 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. En av frågorna som undersökningen vill få svar på 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

Betafördelningen är en kontinuerlig fördelning på $[0,1]$-intervallet med två parameterar $\alpha > 0$ och $\beta > 0$ och skrivs $X \sim Beta(\alpha, \beta)$. Dess täthet ges av

$$f(x) = x^{\alpha-1}(1-x)^{\beta-1}/b(\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \leq x \leq 1,$$

där $b(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ är normaliseringskonstanten som gör att tätheten integrerar till 1 (och är känd som beta-funktionen). Notera att $Beta(1,1)$ är den likformiga fördelningen på $[0,1]$.

Antag att vi vill skatta en sannolikhet p, som alltså ligger i $[0,1]$, t ex sannolikheten att en slumpvis vald person vet vem som vann senaste Eurovision. Låt oss anta en Bayesiansk hållning. Ett neutralt val av apriori-fördelning av p är likformig fördelning eftersom vi tror lika mycket på alla p-värden, dvs $p \sim Beta(1,1)$ (där p spelar rollen som x ovan).

Antag att $n = 10$ slumpvis valda personer tillfrågas om vem som vann senaste Eurovision och att $y = 7$ personer ger rätt svar (=Loreen).

a) Vilken fördelning är $y = 7$ en observation av? Ange likelihooden för detta utfall och härled ML-skattningen för p och dess numeriska värde. (4 p)

b) Antag nu det Bayesianska synsättet med apriorifördelning $p \sim Beta(1,1)$ och härled aposteriofördelning för p givet att $y = 7$. Du behöver bara bry dig om den delen som beror på p, så om du vill kan du stryka ev faktorer som inte innehåller p (om du t ex kommer fram till svaret $71p$ så kan du svara bara $\propto p$ – eftersom det är en täthet ges ju konstanten av att täthetens integral ska bli 1). (2 p)

c) Vilken fördelning har denna aposteriorifördelning? Vad blir motsvarande svar om vi i stället hade valt $Beta(\alpha, \beta)$ för godtyckligt val av $\alpha > 0$ och $\beta > 0$ som apriorifördelning? (4 p)

Lycka till!