STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
TENTAMEN
10 januari 2024

Tentamen i Statistisk analys

10 januari 2024 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 17/1 2024 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.

Tillägg till formelsamlingen: Om observationer sker oberoende med k antal möjliga utfall och sannolikheten för ett utfall i under H_0 är p_i, så kan H_0 testas genom att bilda $Q = \sum_i (n_i - e_i)^2/e_i$, där n_i är antalet observationer som resulterade i utfall i och e_i motsvarande förväntat antal observationer under H_0 (summan kan i vissa fall skrivas som en dubbelsumma).

Under H_0 är $Q \sim \chi^2$ vilket gäller approximativt om alla $e_i \geq 5$. Antalet frihetsgrader är antalet möjliga olika utfall subtraherat med antal icke-fria sådana (beroende på antal skattade parameterar och att totalt antal observationer är givet).

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) Om ett p-värde blir större än 0.05 så leder det till att H_0 *ej* förkastas på 5%-nivån.

b) En normalfördelningsplot är till för att illustrera Centrala gränsvärdessatsen, dvs att summor av slumpvariabler tillsammans blir normalfördelade.

c) Icke-parametriska metoder är speciellt lämpat när stickprov kommer från tjocksvansade fördelningar och när data inte är numeriska.

d) Huvudproblemet vid omfattande bortfall i urvalsundersökningar är inte att man får färre observationer utan att bortfallet oftast avviker systematiskt från den svarande populationen.

e) Om man testar en hypotes på 5%- respektive 1%-nivån så kommer 5%-testet ha större styrka.

Uppgift 2

Följande stickprov samlades in (x_i): 70.4, 68.3, 69.4, 71.0, 71.7, 69.1, 69.9, 68.8. Datamaterialet ger upphov till följande summor: $\sum_{i=1}^{8} x_i = 558.6$ och $\sum_{i=1}^{8} x_i^2 = 39013.56$.

a) Konstruera ett 95% konfidensintervall för stickprovets väntevärde μ. (5 p)

b) Testa på 1%-nivån om $\mu = 71$ mot alternativet att $\mu \neq 71$.

(5 p)

Uppgift 3

Ett företag satsar varje månad reklampengar för att öka sin försäljning av en viss vara (belopp anges i tkr). Under 6 månader satsades följande belopp och följande totala försäljningsbelopp blev resultatet. De olika beloppen är korrigerade för naturliga säsongsvariationer m.m., så observationerna kan betraktas som oberoende. Observationerna var (x_i, y_i): (22, 78), (10, 51), (28, 87), (33, 88), (25, 75), (13, 61). Data ger upphov till följande summor: $\sum_i x_i = 131$, $\sum_i y_i = 440$, $\sum_i x_i^2 = 3251$, $\sum_i y_i^2 = 33344$, $\sum_i x_i y_i = 10234$.

a) Illustrera data och definiera lämplig modell inklusive vilka antaganden som görs. (2 p)

b) Skatta modellens parametrar och skattningarnas standardavvikelser. (4 p)

c) Att extrapolera är inte lämpligt, men gör ändå en skattning av förväntad försäljning man skulle ha en månad helt utan reklam. Gör även ett 95% konfidensintervall för skattningen. (2 p)

d) Hur mycket ökad förväntad försäljningen verkar varje satsad tkr (i reklam) generera i ökad försäljning? Skatta detta belopp och ge även ett 95% konfidensintervall för det. (2 p)

Uppgift 4

Ett försäkringsbolag genomför en säkerhetsöversyn över ett antal kommersiella fastigheter och vill utvärdera effekten av detta. Man jämför skadebeloppen på 8 fastigheter året innan och året efter översynen, med följande utfall (i tkr): $(21.1, 3.1)$, $(9.9, 8.1)$, $(32.8, 21.4)$, $(6.1, 2.8)$, $(15.5, 9.1)$, $(2.1, 4.5)$, $(7.8, 4.8)$, $(14.4, 15.1)$.

Företaget undrar om säkerhetsöversynen har en tydlig effekt att minska skadeloppen, eller om säkerhetsöversynen varken gör till eller från.

a) Formulera en lämplig nollhypotes och ensidig mothypotes och argumentera för varför ett parametriskt test kan vara olämpligt.

(3 p)

b) Genomför lämpligt icke-parametriskt test på 5%-nivån.

(7 p)

Uppgift 5

En yrkesfiskare har under många tidigare år fått vänta på acceptabel fångst ca varannan fisketur. Eftersom hen fiskar på olika ställen varje gång kan fångsterna respektive dag anses oberoende. För att avgöra om fångsten gått ned senasta året gjordes ett antal mätningar på hur många dagar det dröjde till nästa god fångst. Utfallet av 40 mätningar blev $x_1 = 3,\ x_2 = 1,\ x_3 = 1, \ldots, x_{40} = 2$, som sammanfattas med $n_1 = 12$, $n_2 = 14$, $n_3 = 10$, $n_4 = 4$ och $n_5 = 2$. Dvs 12 gånger blev det god fångst dagen efter senaste god fångst, 14 gånger dröjde det 2 dagar till nästa god fångst osv.

En rimlig modell för dessa data under H_0: fiskefångsterna har *inte* förändrats sedan tidigare, är att de har följande fördelning för antal dagar till nästa god fångst $p_X(k) = P(X = k) = p(1-p)^{k-1}$ med $p = 1/2$, dvs för-första-gången fördelningen ($ffg(p = 0.5)$).

a) Argumentera varför denna fördelning är rimlig. Förklara även varför det är viktigt att fiskningarna sker på olika ställen, t ex genom att förklara varför antagandet kanske inte gäller om fiskaren fiskar två dagar i rad på samma ställe innan hen byter fiskeplats. (2 p)

b) Testa H_0: data kommer från $ffg(p = 0.5)$, på 5%-nivån. Detta kan göras genom att jämföra $n_i, \ldots, n_{4+}$ med vad de borde vara under H_0 på lämpligt sätt. För att testet ska fungera på lämpligt sätt bör du slå ihop de sista två utfallen till $n_{4+} = 6$.

(8 p)

Uppgift 6

Använd samma data $x_1, \ldots, x_{40}$ som i uppgift 5 (som kan sammanfattas med $n_1 = 12$, $n_2 = 14$, $n_3 = 10$, $n_4 = 4$ och $n_5 = 2$). Nu bör du ej slå ihop de två sista utfallen.

a) Skriv upp likelihooden som funktion av p för dessa data. Du kan antingen skriva likelihooden för $x_1, \ldots, x_{40}$ eller för de sammanfattade data $n_1, \ldots n_5$ (de är lika så när som på kombinatoriska storheter). (4 p)

b) Härled ML-skattningen $\hat{p}$ för p och ange vad den blir numeriskt. (4 p)

c) Tolka skattningen $\hat{p}$. T ex kan undersökningen även beskrivas med att fiskaren var ute ett visst antal dagar och att 40 av dessa dagar resulterade i god fångst. Vad blir blir ML-skattningen uttryckt i dessa storheter? (2 p)

Lycka till!