

Tentamenskrivning: TMS145 - Grundkurs i matematisk statistik och bioinformatik, 7,5p.

Tid: Onsdagen den 3 april 2013, kl 08.30 - 12.30.

Examinator: Erik Kristiansson

Jour: Erik Kristiansson, tel 070-5259751

Hjälpmedel: kalkylator, egen handskriven formelsamling (fyra A4 sidor) samt med skrivningen utdelade tabellsidor.

Max är 32 poäng. För godkänt krävs minst 15 poäng, för betyget 4 krävs 20 poäng och för 5 krävs 25 poäng. Uppgifterna kommer inte i svårighetsordning.

1. Månadslönen före skatt hos en heltidsarbetande slumpmässigt utvald svensk medborgare kan beskriva med hjälp av en Paretofördelad stokastisk variabel X som har täthetsfunktionen

$$f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}, \quad x \geq \beta.$$

Antag att $\alpha = 2$ och $\beta = 15000$.

- (a) Beräkna sannolikheten att lönen före skatt hos en heltidsarbetande slumpmässigt utvald svensk medborgare är mellan 30000 och 60000? Vad är sannolikheten att den överstiger 60000?
- (b) Visa att den förväntade lönen $E[X]$ är större än medianlönen m_X .

(5 p)

2. Koncentrationen av DNA i ett prov mäts med hjälp av en spektrofotometer. Eftersom mätningen är osäker upprepas den 10 gånger $(x_1, \dots, x_{10})$ där medelvärdet beräknas till $\bar{x} = 4.87$ ($\mu\text{g/mL}$). Antag att observationerna från mätningen är oberoende och slumpmässigt dragna från en normalfördelning med okänt väntevärde μ och känd varians $\sigma^2 = 4$.

- (a) Beräkna ett dubbelsidigt konfidensintervall för väntevärdet av mängden DNA (μ). Konfidensgraden ska vara 99%.
- (b) Om du ändrar stickprovsstorleken från 10, hur många mätningar måste göras om längden på konfidensintervallet inte ska överstiga 1 ($\mu\text{g/mL}$)?

(4 p)

3. Låt X beteckna antalet poäng en slumpmässig bioteknikstudent får på den inledande uppgiften i en tentamen i kursen TMS145. Antag att X har följande sannolikhetsfördelning

$$p_X(x) = \begin{cases} 0.15 & x = 0 \\ 0.05 & x = 1 \\ 0.20 & x = 2 \\ 0.35 & x = 3 \\ 0.25 & x = 4. \end{cases}$$

Antag vidare att poängen från olika studenter är oberoende och likafördelade. Låt $\bar{X}$ vara medelpoängen av 52 tenterande bioteknikstudenter.

- (a) Beräkna väntevärdet och variansen för X .
- (b) Använd centrala gränsvärdessatsen för att beräkna en approximativ fördelning för $\bar{X}$. Glöm inte att motivera approximationen.
- (c) Beräkna den approximativa sannolikheten att medelpoängen på uppgiften bland de 52 tenterande är större än 3.

(5 p)

4. Efter ett miljöutsläpp av stora mängder antibiotika undersöks förekomsten av motståndskraftiga bakterier. Därför samlades två stickprov in, ett med 35 jordprov från den förorenade miljön ($x_1, \dots, x_{35}$) och ett med 43 jordprov från en kontrollmiljö ($y_1, \dots, y_{43}$). Mängden motståndskraftiga bakterier uppmättes sedan med hjälp av PCR och följande medelvärden och standardavvikelser beräknades $\bar{x} = 0.110$, $s_x = 0.109$, $\bar{y} = 0.0479$ och $s_y = 0.011$ (mätvärdena beskriver relativ mängd motståndskraftiga bakterier). Av tidigare erfarenhet vet man att variationen i de olika stickproven ej kan antas vara lika.

- (a) Formulera lämpliga hypoteser och fördelningsantaganden. Genomför sedan ett enkelsidigt test för att undersöka om mängden motståndskraftiga bakterier är högre i den förorenade miljön. Signifikansnivån ska vara 0.01.
- (b) Beräkna p-värdet för testet i (a).

(5 p)

5. Låt X vara en stokastisk variabel med täthetsfunktionen

$$f_X(x) = (a + 1)x^a, \quad 0 < x < 1,$$

där a är en okänd positiv parameter.

- (a) Använd momentmetoden för att härleda en punktskattning $\hat{a}_M$ av a .
- (b) Använd maximum-likelihoodmetoden för att härleda en punktskattning $\hat{a}_{ML}$ av a .
- (c) Antag att $X_1, \dots, X_5$ antar värdena 0.87, 0.93, 0.50, 0.66, 0.80. Beräkna observationsvärden på $\hat{a}_M$ och $\hat{a}_{ML}$. Är punktskattningarna lika? Förklara.

(5 p)

6. Låt X och Y vara två kontinuerliga och oberoende stokastiska variabler som båda är likformigt fördelade mellan 0 och 1.

- (a) Beräkna $P(X + Y \leq 0.5)$.
- (b) Låt $Z = -\ln(X)$. Visa att Z är exponentialfördelad med parameter $\lambda = 1$.

(4 p)

7. (a) What is the Levenshtein distance between strings "ATA" and "TAA"? Explain your answer.
- (b) Assuming a match score of 2, a mismatch score of -1 and a gap score of -2, derive the score matrix for a global alignment of "ATA" and "TAA".

In this case, what is the score of an optimal global alignment? How many alignments have this optimal score (remember: each path represents a different alignment)? What are these alignments?

(4 p)