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

Tid: Onsdag den 19 augusti 2009, kl 08:30-12:30 Väg och vatten
Tesen korrigerad enligt anvisnigar under tentamenstillfället.
Examinator: Olle Nerman, tel 7723565.
Jour: Frank Eriksson, tel 073-7263051.
Hjälpmedel: Egen handskriven formelsamling (fyra A4-sidor) samt med skrivningen utdelade tabellsidor.

Maxpoäng: 32. För godkänt krävs minst 15 poäng totalt och minst 4 poäng på sannolikhetsteori- och statistik-delen vardera samt minst 3 poäng på bioinformatikdelen. För betyget 4 krävs 20 poäng och för betyget 5 25 poäng.

Sannolikhetsteori

1. a Vad menas med medianen för en kontinuerligt fördelad stokastisk variabel?

   (2p)

   b Vad är medianen för en exponentialfördelad stokastisk variabel med väntevärde 3?

   (2p)

2. a Beräkna sannolikheten vid kast av 3 vanliga tärningar (med poängsidorna 1,2,3,4,5 eller 6) för händelsen A = *"summan av kasten blir exakt 5"*?

   (2p)

   b Vad är den betingade sannolikheten vid kast av 3 vanliga tärningar för händelsen B=" *minst en tärning visar poängen 3*" givet händelsen A=" *summan av kasten är exakt 5*"?

   (2p)

Ledning: Räkna möjliga och gynnsamma fall

Vänd!

3. Ett visst flygplan kan på grund av säkerhetsskäl, när det är fullsatt, ha en obalans mellan stolsraderna på vänster och högersidan som är maximalt 250KG. På varje sida finns 50 platser. En typisk passagerare har en vikt som har väntevärdet 80 Kg och standardavvikelsen 10 Kg. Antag oberoende mellan resenärernas vikt på resp. plats och att planet är fullsatt.

   a Vilket väntevärde har totala viktskillnaden (med tecken) mellan passagerarna till höger och vänster i ett fullsatt plan?

   (1p)

   b Vilken varians har totala viktskillnaden mellan passagerarna till höger och vänster i ett fullsatt plan?

   (1p)

   c Vad är approximativt sannolikheten för att obalansen överskrider gränsen 250 kg.

   (2p)

Statistik

4. a Tag fram Maximum Likelihood-skattaren för θ ur ett stickprov från en Rayleigh-fördelning, dvs. en fördelning med täthetsfunktion

$$f(x) = \theta x e^{-\theta x^2/2}, \quad x \geq 0.$$

   (2p)

   b Ange två önskvärda egenskaper hos en punktskattare.

   (2p)

5. a Vilken är fördelen med att använda intervallskattning istället för punktskattning?

   (1p)

   b Ge definitionen av *p-värde*.

   (1p)

   c Anatg att du har ett stickprov av storlek n från en normalfördelning med väntevärde μ och varians $\sigma^2 = 4$. Du vill testa $H_0 = \mu = 10$ mot $H_1 : \mu < 10$ på signifikansnivå 0.01. Du använder den vanliga teststatistikan, d.v.s. $(\overline{X} - \mu)/(\sigma/\sqrt{n})$. Antag att det sanna väntevärdet är 9. Hur stor bör stickprovsstorleken n vara för att teststyrkan (power) blir minst 90%?

   (2p)

Vänd!

6. Vikt X (pounds) och systoliskt blodtryck Y (blodtryck vid hjärtats sammandragning) hos 26 slumpmässigt utvalda (och oberoende) män i åldrarna 25-30 visas i tabellen nedan.

Patientnr	Vikt	Blodtryck	Patientnr	Vikt	Blodtryck
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159

Antag nu att vikt och blodtryck är bivariat normalfördelade. Från värdena ovan får vi att

$$\sum_i x_i = 4743 \quad \sum_i y_i = 3786$$

$$\sum_i x_i^2 = 880545 \quad \sum_i y_i^2 = 555802 \quad \sum_i x_i y_i = 697076$$

$$\sum_i (x_i - \overline{x})^2 \approx 15312 \quad \sum_i (y_i - \overline{y})^2 \approx 4502 \quad \sum_i (x_i - \overline{x})(y_i - \overline{y}) \approx 6422$$

Vänd!

En (något beskuren) utskrift från programmet R med datamänden ovan ger

```
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value
(Intercept) 69.10437   12.91013   5.353
x            0.41942    0.07015   5.979

Residual standard error: 8.681 on 24 degrees of freedom
Multiple R-squared: 0.5983,     Adjusted R-squared: 0.5815
F-statistic: 35.74 on 1 and 24 DF,  p-value: 3.591e-06
---

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq
x          1 2693.58 2693.58
Residuals 24 1808.57   75.36
```

a Ange uttrycket för en linjär regressionsmodell med blodtryck som svarsvariabel baserad på informationen ovan (där alla koefficienter samt regressionslinjens slutliga form anges).

(1p)

b Tolka koefficienterna (uttryckt i termer relaterade till problemet).

(1p)

c Baserat på informationen ovan, finns det ett statistiskt säkerställt samband mellan vikt och blodtryck? Motivera!

(1p)

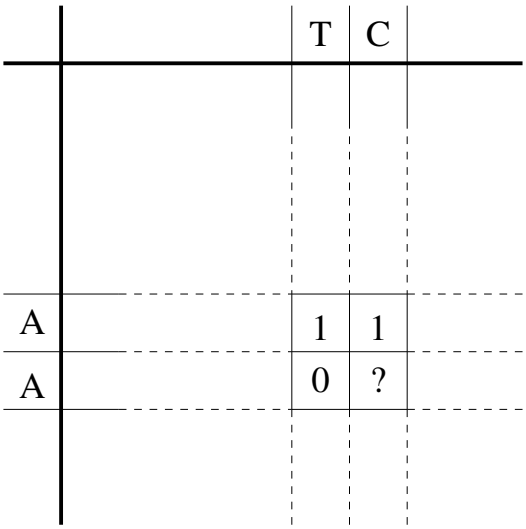
d Hur stor andel av variationen i männens blodtryck kan enligt modellen förklaras av personernas varierande vikt?

(1p)

Vänd!

7. Sekvensbioinformatik

   a The figure below shows part of the score matrix for a local alignment of two sequences.



   When using the Smith-Waterman algorithm for pairwise local alignment, explain how the score for cell (i,j) of the score matrix, usually denoted $F(i, j)$, is calculated.

   Assuming a match score of 3, a mismatch score of -2 and a gap score of -3, explain how the value for the cell containing a question mark is calculated. What value will be placed in that cell of the score matrix?

   (3p)

   b If we want to perform a local alignment of a query sequence with the contents of a very large database of sequences, we could use the Smith-Waterman algorithm or we could use the BLAST program. Give one advantage and one disadvantage of using BLAST.

   (1p)

8. Strukturbioinformatik.

   a What is the purpose of the DSSP program? Describe how main chain hydrogen bonds are calculated by DSSP.

   (2p)

   b Describe the family and superfamily levels in the Structural Classification of Proteins (SCOP).

   (2p)