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

Tid: Onsdag den 18 augusti, 2010 08:30-12:30, Väg och vatten.
Examinator: Olle Nerman, tel 7723565.
Jour: Alexandra Jauhiainen, tel 0737168778
Hjälpmedel: valfri miniräknare, 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 statistikdelen vardera samt minst 3 poäng på bioinformatikdelen.

Sannolikhetsteori

1. a) Ur en kortlek på 52 kort dras på måfå fem kort. Beräkna sannolikheten att man får en flush, dvs fem kort i samma färg.

2p

b) En väl blandad kortlek med 52 kort delas i fyra lika stora delar. Beräkna sannolikheten för att varje del innehåller en kung.

2p

2. Vi har en stokastisk variabel Y med täthetsfunktion

$$f_Y(y) = \begin{cases} \frac{1}{2y\ln(2)} & \text{för } \frac{1}{2} < y < 2, \\ 0 & \text{för övrigt.} \end{cases}$$

a) Beräkna $P(3/4 < Y < 3/2)$.

2p

b) Beräkna väntevärdet för Y.

1p

c) Beräkna standardavvikelsen för Y.

1p

Vänd!

3. a) I ett mycket stort register med DNA-sekvenser förekommer två mönster A och B enligt följande:

   - 8 % av alla sekvenser har både A och B
   - 12 % av alla sekvenser har A men inte B
   - 22 % av alla sekvenser har B men inte A

   Avgör om förekomst av mönster A respektive B är oberoende händelser.

   2p

   b) De stokastiska variablerna X och Y är normalfördelade och oberoende. Vi har $\mu_X = 1$ och $\sigma_X = 1$ samt $\mu_Y = -1$ och $\sigma_Y = 2$. Beräkna sannolikheten att produkten av X och Y är negativ.

   2p

   Observera att det inte finns något samband mellan *a*) och *b*).

Statistik

4. a) Definiera begreppen Likelihood-funktion och Maximum Likelihood-skattare för en parameter θ.

   1p

   b) Låt X vara en stokastisk variabel och $\mu_X = E[X]$. Låt $X_1, \ldots, X_{10}$ vara ett stickprov från fördelningen för X. Betrakta följande två skattare av μ_X

   * $W_1(X_1, \ldots, X_{10}) = \frac{1}{10} \sum_{i=1}^{10} X_i$
   * $W_2(X_1, \ldots, X_{10}) = X_1$

   Vi vet att W_1 är väntevärdesriktig.

   i) Är W_2 en väntevärdesriktig skattare av μ_X? Motivera.

   1p

   ii) Vilken av W_1 och W_2 är att föredra? Motivera.

   1p

Vänd!

5. Vi har samlat in 72 prov av en förorenad jordmån (400g var) som vi har torkat och analyserat för cyanid. Medelcyanidnivån i vårt stickprov är $\overline{x} = 116$ mg/kg och standardavvikelsen $s = 80$ mg/kg.

   a) Testa hypotesen att den sanna cyanidnivån i jordmånen är högre än 100 mg/kg. Använd signifikansnivån 0.1.

   2p

   b) Skulle du dra samma slutsats som i a) om signifikansnivån var 0.05? Eller 0.01? Varför kan signifikansnivån leda till olika beslut?

   1p

   c) Förklara hur du hade kunnat svara på frågan i a) med ett konfidensintervall (vilket?) istället för ett hypotestest. Du behöver inte räkna ut intervallet.

   2p

6. För att undersöka om mängden C-vitamin i frukt minskar med förvaringstiden mättes halten C-vitamin hos ett parti kiwkifrukter vid olika tidpunkter.

Förvaringstid (dagar)	**x**	0	1	2	5	7	10
C-vitaminhalt (mg/100g)	**y**	93	85	80	84	83	79

$$\bar{x} = 4.17,\ \bar{y} = 84,\ s_{xx} = 74.83,\ s_{xy} = -64,\ s_{yy} = 124$$

   a) Sätt upp en linjär regressionsmodell och skatta samtliga parametrar i modellen. Vilka antaganden görs?

   2p

   b) Kan vi baserat på våra data dra slutsatsen att C-vitaminhalten i kiwifrukt minskar linjärt med förvaringstiden (i det aktuella tidintervallet)?

   2p

Vänd!

Bioinformatik

7. Sequence Alignment

   Using a gap score of -2 and match/mismatch scores taken from the PAM250 substitution matrix (given below), derive the score matrix for a global alignment of QFN with $NGYE$.

   In this case, what is the score of an optimal global alignment? Give the alignment(s) with this score.

   PAM250 substitution matrix:

```
    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A   2
R  -2  6
N   0  0  2
D   0 -1  2  4
C  -2 -4 -4 -5  4
Q   0  1  1  2 -5  4
E   0 -1  1  3 -5  2  4
G   1 -3  0  1 -3 -1  0  5
H  -1  2  2  1 -3  3  1 -2  6
I  -1 -2 -2 -2 -2 -2 -2 -3 -2  5
L  -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6
K  -1  3  1  0 -5  1  0 -2  0 -2 -3  5
M  -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6
F  -4 -4 -4 -6 -4 -5 -5 -5 -2  1  2 -5  0  9
P   1  0 -1 -1 -3  0 -1 -1  0 -2 -3 -1 -2 -5  6
S   1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  3
T   1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -2  0  1  3
W  -6  2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4  0 -6 -2 -5 17
Y  -3 -4 -2 -4  0 -4 -4 -5  0 -1 -1 -4 -2  7 -5 -3 -3  0 10
V   0 -2 -2 -2 -2 -2 -2 -1 -2  4  2 -2  2 -1 -1 -1  0 -6 -2  4
```

4p

Vänd!

8. Structural Bioinformatics

a) Describe the heuristics that can be used in predicting protein secondary structure manually from a multiple sequence alignment.

2p

b) In the CATH classification of protein domain structures, the letters in the name $CATH$ represent the four major levels in the classification hierarchy. What are the names of the four major levels in CATH? Describe the levels represented by the letters A and T.

2p