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

Tid: Tisdag den 18 december, 2007 kl 8.30 - 12.30 i V-huset.
Examinator: Olle Nerman, tel 7723565.
Jour: Alexandra Jauhiainen, tel 073-7168778, Erik Kristiansson, tel 070-5259751.
Hjälpmedel: valfri miniräknare, egen handskriven formelsamling (fyra A4 sidor) samt med skrivningen utdelade formel- och 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.

Sannolikhetsteori

1. Skevheten för en stokastisk variabel X är ett mått på asymmetrin i dess fördelning. Skevheten betecknas med γ och kan beräknas enligt formeln
$$\gamma = \frac{\mathbb{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3},$$
där μ är väntevärdet och σ är standardavvikelsen för X.

   (a) Låt Y_1 vara likformigt fördelad på intervallet (-1, 1), dvs med parametrar $a = -1$ och $b = 1$. Beräkna skevheten för Y_1. (1p)

   (b) Låt Y_2 vara fördelad enligt täthetsfunktionen
$$f_{Y_2}(y) = \frac{1}{2}(1 + y), \quad -1 < y < 1.$$
Beräkna skevheten för Y_2. (3p)

2. (a) Förra påsken skrev 10 bioteknikstudenter omtentan i matematisk statistik. Av dessa fick 6 godkänt och de övriga 4 fick underkänt. Antag att vi väljer 5 av de 10 studenterna slumpmässigt. Vad är då sannolikheten att minst 4 av de 5 utvalda fick godkänt på omtentan? (2p)

   (b) Antag att sannolikheten att få godkänt på omtentan i matematisk statistik är 0.60. Om 1.000 studenter skriver tentan, vad är då approximativt sannolikheten att minst 625 av dessa får godkänt? (2p)

3. Låt X och Y vara livslängden (i år) för två komponenter och antag att de är fördelade enligt den simultana täthetsfunktionen

$$f_{X,Y}(x,y) = e^{-x}, \quad 0 < y < x.$$

   (a) Visa att X och Y ej är oberoende. (2p)

   (b) Beräkna den betingade sannolikheten att $X > 1$ givet att $Y > 1$. (2p)

Statistik

4. Man gör undersökningar vid badplatser för att få reda på om halten av gift från giftalger överstiger vissa gränsvärden. Om man kan visa att den förväntade nivån överstiger 0.6 sätter man upp en varningsskylt som rekommenderar folk att ej bada där och om man kan visa att den förväntade nivån överstiger 0.8 utfärdar man badförbud. Man observerar nivåerna

   ```
   1.18  0.99   0.83   0.71   1.27   0.49   1.58   1.05
   ```

   Antag att nivåernas värden är oberoende och varierar enligt normalfördelning.

   Följande info kan vara användbar: stickprovsmedelvärdet=1.0125 och stickprovsstandardavvikelsen=0.341.

   (a) Ska man sätta upp varningsskylten? Genomför lämplig undersökning på nivå 0.01. (2p)

   (b) Ska man utfärda badförbud? Genomför lämplig undersökning på nivå 0.01. (2p)

5. (a) En klass med 320 juridikstudenter delades slumpmässigt in i två lika stora grupper. Båda grupperna fick sedan lyssna på en inspelning av en rättegång i ett rattfyllerifall. Grupp 1 fick sedan se en bild på en välklädd person med trevligt utseende som påstods vara den åtalade. Grupp 2 fick istället se en bild på en illa klädd person med otrevlig uppsyn som påstods vara den åtalade.

   Därefter fick studenterna oberoende av varandra välja ett av besluten 'skyldig' och 'icke-skyldig'. I grupp 1 valde 72 studenter 'skyldig' medan 91 studenter valde samma alternativ i grupp 2.

   Tyder detta på att utseendet kan ha betydelse för domslutet? Motivera ditt svar med ett lämpligt konfidensintervall med konfidensgrad 0.95 eller test på nivå 0.05. Skriv tydligt din slutsats. (2p)

(b) Man har studerat avståndet i meter mellan målet och nedslagsplatsen för en missil. Vid 60 provskjutningar fick man förljande resultat:

Avstånd	0-100	100-200	200-300	300-400
Frekvens	25	17	11	7

Man hävdar att avståndet mellan målet och nedslagsplatsen är likformigt fördelat med parametrarna 0 och 400. Pröva detta påstående på lämpligt sätt på nivån 0.01. (2p)

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

6. (a) Låt $x_1, \ldots, x_n$ vara observerade kölängder i ett kösystem. Vi antar att observationerna kommer från oberoende stokastiska variabler $X_1, \ldots, X_n$. Sannolikhetsfunktionen för X_i är

$$p(x) = \theta^x(1-\theta) \quad \text{för} \quad x = 0, 1, 2, \ldots$$

där θ är okänd.

Ta fram maximum likelihood-skattningen av θ. (2p)

(b) I regressionsanalys arbetar man ofta med kvadratsummor. Beskriv de intressanta kvadratsummor som man tittar på och vilken information man kan få från dessa. (2p)

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

Bioinformatik

7. Sekvensbioinformatik

(a) 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 "GAATA" with "GATAC".

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)? (3p)

(b) Calculate the score of the following multiple alignment using the BLOSUM62 matrix and the sum of pairs method:

Sequence 1: HCDV
Sequence 2: HAHV
Sequence 3: HATA

(1p)

BLOSUM62 Matrix:

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  4
R -1  5
N -2  0  6
D -2 -2  1  6
C  0 -3 -3 -3  9
Q -1  1  0  0 -3  5
E -1  0  0  2 -4  2  5
G  0 -2  0 -1 -3 -2 -2  6
H -2  0  1 -1 -3  0  0 -2  8
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
```

8. Strukturbioinformatik.

(a) Draw a sketch of a Ramachandran plot. Explain what a Ramachandran plot shows. (2p)

(b) What are the main steps in the comparative modelling process? (2p)