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

Tid: Lördag den 29 mars, 2008 kl 14.00 - 18.00 i V-huset.
Examinator: Olle Nerman, tel 7723565.
Jour: Alexandra Jauhiainen, tel 073-7168778.
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. Anropen till en viss telefonstation antas komma som en Poissonprocess med 5 anrop/minut, dvs $X \sim \text{Poi}(5t)$, där X är antalet anrop under ett t minuter långt tidsintervall.

   (a) Vad är chansen att det under ett visst 2 minuters-intervall kommer precis 10 samtalsanrop? (2p)

   (b) Approximera sannolikheten för minst 60 samtal under ett visst 10-minutersintervall. (2p)

2. Antag att du har ett övergångsställe där det växlar mellan rött och grönt för gående med 30 sekunders intervall (rött ljus 30 s. och grönt ljus 30 s.).

   (a) Hur ser kumulativa fördelningsfunktionen ut för din väntetid X vid en "slumpmässig" ankomst till övergångsstället? (2p)

   För en positiv stokastisk varabel Y som är begränsad gäller att väntevärdet kan definieras som

   $$\int_0^\infty P(Y > x)dx$$

   (b) Beräkna väntevärdet $E[X]$ för X (definierad som i a-uppgiften) (1p)

   (c) Beräkna variansen för X (definierad som i a-uppgiften) (1p)

3. Antag att du upprepar ett försök n gånger (oberoende försöksupprepningar) och räknar antalet gånger X som en viss händelse A i försöket inträffar. På samma sätt räknar du också ut Y antalet gånger som en annan händelse B inträffar.

   (a) Uttryck kovariansen för X och Y som en funktion av n, $P(A)$, $P(B)$ och $P(A \cap B)$. (2p)

   (b) Beräkna variansen för $Y - X$ uttryckt i n, $P(A)$, $P(B)$ och $P(A \cap B)$. (1p)

   (c) Visa att om A och B är oberoende händelser så är X och Y oberoende stokastiska variabler. (1p)

Statistik

4. En viss typ av transistorer har exponentialfördelad livslängd. Man sätter 400 transistorer av denna typ i bruk samtidigt och konstaterar efter 1 tidsenhet att endast 109 fungerar.

   (a) Skatta medellivslängden. (2p)

   (b) Skatta medianlivslängden. (2p)

5. I en trafiksäkerhetsstudie noterade man för 42 städer i USA dels $x=$ andelen körkortsinnehavare under 21 år (enhet procent), dels $y=$antalet dödsolyckor per 1000 körkortsinnehavare och år med följande resultat:

x	y	x	y
8	0.885	12	2.246
8	0.368	12	1.913
8	0.645	13	2.962
8	2.190	13	1.142
8	0.820	13	2.634
8	1.267	14	2.855
9	1.028	14	2.352
9	1.433	14	2.890
9	0.338	14	1.443
9	0.835	14	1.643
9	0.926	15	2.623
10	0.039	15	3.224
10	1.014	15	2.814
10	0.493	16	2.801
10	1.926	16	3.623
11	2.091	16	2.943
11	1.849	17	2.627
11	1.294	17	4.100
12	0.708	17	3.256
12	1.652	18	3.830
12	1.405	18	3.614

Analysresultat från R:

```
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-1.234118 -0.264408  0.007723  0.443619  1.490989

Coefficients:
            Estimate      Std. Error
(Intercept) -1.59741      0.37167
 x           0.28705      0.02939
---
Analysis of Variance Table

Response: y
          Df    Sum Sq
x          1    33.134
Residuals 40    13.893
```

(a) Förefaller andelen körkortsinnehavare under 21 år att vara av betydelse för antalet dödsolyckor? Besvara frågan med ett lämpligt test på nivå 0.001 eller ett lämpligt konfidensintervall med konfidensgrad 0.999. (2p)

(b) I en "pensionärsstad" finns 5% körkortsinnehavare under 21 år. Kan man använda modellen ovan för att prediktera antalet dödsolyckor i denna stad? Motivera kortfattat. (1p)

(c) Ge uttrycket för förklaringsgraden i en regressionsmodell samt ange förklaringsgraden i modellen ovan. (1p)

6. För ett slumpmässigt stickprov med tio observationer, $x_1, \ldots, x_{10}$ från $X_1, \ldots, X_{10}$ där $X_i \sim N(\mu, \sigma^2)$, har man beräknat stickprovsstandardavvikelsen och fått $s = 3.21$.

   (a) Pröva $H_0 : \sigma = 2.5$ mot $H_1 : \sigma > 2.5$ på nivån 0.05. (2p)

   (b) Bestäm det σ-värde för vilket testet har styrkan 0.90. (2p)

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 local alignment of "GAAC" with "TGAC".

   In this case, what is the score of an optimal local alignment? How many alignments have this optimal score (remember: each path represents a different alignment)? What are these alignments? (2p)

   (b) The PAM250 matrix is shown below. Comment on the scores between W and W; A and A; I and L; F and D. (2p)

	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

8. Strukturbioinformatik.

   a) Describe the "Architecture" and "Topology" levels in the CATH protein structure classification. Draw sketches of two folds (these do not have to be real protein folds) that have the same architecture (as defined in CATH), but have different topologies. (2p)

   (b) By drawing a sketch of a proline residue and a sketch of a Ramachandran plot, explain how the constraints on the main-chain conformation of proline residues differ from those of other amino acid residues. (2p)