

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

Tid: Tisdag den 15 december, 2009 kl 8.30 - 12.30, Hörsalsvägen.

Examinator: Olle Nerman, tel 7723565.

Jour: Frank Eriksson, tel 0737263051

Hjälpmedel: valfri miniräknare, egen handskrivna 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 sannolikhetssteori- och statistikdelen vardera samt minst 3 poäng på bioinformatikdelen.

Sannolikhetssteori

1. Tre nukleotider vardera av A , C , T , och G ska länkas samman till en DNA-sekvens.

(a) Hur många sekvenser med olika utseenden kan bildas?

(2p)

(b) Antag att alla konfigurationer i (a) är lika troliga. Hur stor är sannolikheten att alla tre nukleotiderna av vardera sort ligger intill varandra, som t.ex. i fallet $AAATTTCCCGGG$?

(2p)

2. Antag att avstånd mellan bilar på en landsväg (i en riktning) kan antas vara oberoende och exponentialfördelade med väntevärde 100 (meter). Tänk dig att du befinner dig i en bil på denna väg.

(a) Vilken (approximativ) fördelning har avståndet till den femtionde bilen framför dig?

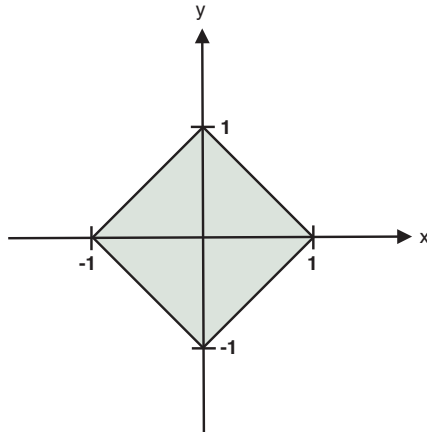
(2p)

(b) Vad är sannolikheten att avståndet till den femtionde bilen är mellan 5 och 5.2 km?

(2p)

Vänd!

3. Antag att vi har en tvådimensionell stokastisk variabel (X, Y) med simultan täthetsfunktion $f_{X,Y}(x, y) = 1/2$ för $|x| + |y| \leq 1$. Det intressanta området är skuggat i nedanstående figur.



- (a) Bestäm de marginella täthetsfunktionerna för X och Y .

(2p)

- (b) Bestäm väntevärdena $E[X]$ och $E[Y]$.

(1p)

- (c) Är X och Y oberoende? Motivera!

(1p)

Vänd!

Statistik

4. Antag att X som är Binomialfördelad med parametrar (n, p) där n är känd men p okänd.

(a) Finn maximum likelihoodskattaren av parametern p baserat på en ensam observation av X .

(2p)

(b) Finn maximum likelihoodskattaren av variansen $Var(X)$ baserat på en ensam observation av X .

(1p)

(c) Vad menas med att en skattare är väntevärdesriktig (unbiased)?

Är skattaren av p i (a) väntevärdesriktig?

Är skattaren av $Var(X)$ i (b) väntevärdesriktig?

Ledning: För varje stokastisk variabel Z gäller att

$$Var(Z) = E[(Z - E[Z])^2] = E[Z^2] - (E[Z])^2.$$

(2p)

5. Det är känt att nativiteten (fördelsetalet) inte är jämnt utspritt över året. Nativiteten skiljer sig mellan olika månader och i allmänhet föds fler barn under våren och sommaren än under vinterhalvåret.

Men hur är det med mortaliteten (dödstalet)? Är den jämnt utspridd över året? Sätt upp lämpligt test och testa detta, baserat på 40 slumpmässigt valda dödsattester. När du granskar de 40 dödsattesterna ser du att av dessa personer dog 9 personer under det första kvartalet, 12 under andra, 12 under tredje och 7 under fjärde. Vilken är din slutsats?

(3p)

Vänd!

6. För drygt hundra år sedan demonstrerade den danske näringsforskaren Mikkel Hindhede att det går att överleva på enbart potatis. Detta gjorde han genom att bestämma att laboratoriebiträdet Fredrik Madsen skulle leva på potäter allena under nästan ett års tid. Det dagliga potatisintaget uppgick till nästan 4 kg.

Tyvärr saknar vi mätningar från Hindhedes försök. I ett mer nutida försök mättes askorbinsyrahalten i konventionellt odlad bintjepotatis med ekologiskt odlad King Edward i syfte att försöka påvisa att den ekologiskt odlade potatisen är nyttigare (askorbinsyrahalten antas ge en god uppfattning om C-vitaminnivån). Tabellen nedan sammanfattar mätningarna (enheten är mg/100 g). Potatisarna kan antas oberoende av varandra.

	Ekologisk, X	Konventionell, Y
<i>Stickprovsmedel</i>	$\bar{x} = 14.49$	$\bar{y} = 12.74$
<i>Stickprovsstorlek</i>	$n_X = 38$	$n_Y = 34$
<i>Stickprovsvarians</i>	$s_X^2 = 13.18$	$s_Y^2 = 9.33$

- (a) Undersök genom att konstruera lämpligt enkelsidigt konfidensintervall med konfidsgrad 0.95 om dessa observationer styrker forskarnas tes att den ekologiskt odlade potatisen innehåller en högre halt askorbinsyra.

(2p)

- (b) Vilket hypotestest motsvarar konfidensintervallet i (a)? Beskriv testet genom att ange nollhypotes, mothypotes samt nivå. Svara utan att utföra själva testet på vilken slutsats testet påbjuder.

(1p)

- (c) Definiera begreppen *nivå* samt *styrka* i samband med hypotestest och förklara varför dessa är intressanta.

(1p)

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 *GFQW* with *GNW*.

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) Suppose that $hbond(a, b)$ means that there is a hydrogen bond between the main chains of residues a and b , and that residue a is the hydrogen bond donor and residue b is the hydrogen bond acceptor. In identifying an antiparallel bridge (between residues in antiparallel strands), the DSSP program looks for the following patterns of main chain hydrogen bonds:

$$[hbond(i, j) \text{ and } hbond(j, i)]$$

or

$$[hbond(i - 1, j + 1) \text{ and } hbond(j - 1, i + 1)].$$

For the second of these two cases, draw a sketch to illustrate the hydrogen bond pattern. You should draw the main chain of residues $i - 1, i, i + 1, j - 1, j$ and $j + 1$ (showing atoms N, CA, C and O for each residue). Identify all atoms in your sketch, and use subscripts to identify the residue to which each atom belongs. Show the hydrogen bonds with dotted lines.

(2p)

- (b) When using a comparative modelling approach, why might it be necessary to use a fragment-fitting approach to remodel a loop? Why could it be useful to cluster the candidate conformations for the replacement loop?

(2p)