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

Tid: Lördag den 14 april, 2007 kl 14.00 - 18.00 i V-huset.
Examinator: Olle Nerman, tel 7723565.
Jour: Alexandra Jauhiainen, tel 073-7168778, Erik Kristiansson, tel 7725342.
Hjälpmedel: kalkylator, 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 I en urna finns två röda och två svarta bollar. En person drar på måfå en boll i taget utan återläggning. Låt X vara antalet dragningar tills båda de röda bollarna är dragna.

   (a) Beräkna sannolikhetsfunktionen $p_X(x)$. (2p)

   (b) Beräkna väntevärdet för X. (1p)

   (c) Beräkna standardavvikelsen för X. (1p)

2 En planka med längden 1 meter kapas på en slumpmässigt position (varje position är lika sannolik). Låt X vara längden av den vänstra biten. Denna bit kapas igen (på samma sätt som tidigare). Låt Y vara längden av den vänstra biten efter andra kapningen.

   (a) Beräkna den tvådimensionella täthetsfunktionen $f_{X,Y}(x,y)$ och skissa definitionsområdet. (1p)
   Tips: $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$

   (b) Beräkna väntevärdet för Y. (1p)

   (c) Beräkna kovariansen mellan X och Y. Är X och Y oberoende? (2p)

3 Låt X vara tiden det tar för en räkneövningsledare i kursen "Introduktion till matematisk statistik och bioinformatik för Bt 2" att hjälpa en elev med ett räkneproblem. Antag att X har täthetsfunktionen

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda},\ x \geq 0$$

med parametern $\lambda = 3$ minuter. Antag att tiden det tar att hjälpa varje elev är oberoende och låt Y vara den tiden det tar att hjälpa 30 elever.

(a) Beräkna en approximativ fördelning för Y. (2p)

(b) Examinatorn för kursen vill att räkneövningsledaren ska vara sysselsatt hela räkneövningen samtidigt som den inte får dra över tiden. Om vi antar att det är 30 elever som vill ha hjälp under en räkneövningspass, vad är då sannolikheten att Y avviker med mer än 10 minuter från den totala räkneövningstiden (som är 90 minuter)? (2p)

Statistik

4 Toxaphene är en numera förbjuden organisk pesticid som började användas när DDT förbjöds. Vid inandning eller intag kan lungor, njurar och nervsystemet skadas, ibland fatalt. I fisk orsakar toxaphen C-vitaminbrist som påverkar tillväxt hos fisken och ger upphov till benskörhet.

Ett stort antal individer av två arter sötvattenfiskar odlades i närvaro av samma mängd toxaphene. Ur de två grupperna togs ett stickprov om 200 fiskar från art A och 300 fiskar från art B. Inom art A hade 30 fiskar ryggmärgsskador, medan 70 av fiskarna hade det i art B.

(a) Punktskatta andelen fiskar i varje art som har ryggmärgsskador med momentmetoden. Undersök om skattningarna är väntevärdesriktiga. (2p)

(b) Undersök på lämpligt sätt om de teoretiska frekvenserna av ryggmärgsskador kan anses vara lika i de båda arterna. Nivå 0.05. (2p)

5 Mängden kvarvarande kolmonoxid (CO) (y) mätt i en persons lungor efter det att personen rökt en cigarett är sammanställd i nedanstående tabell tillsammans med tiden sedan den sista cigaretten röktes (x).

x: tid (timmar)	y: CO (ppm)
0.50	53
1.50	22
2.00	38
6.00	17
2.25	28
1.50	32
1.25	35
0.75	40
0.15	61
2.00	22
3.15	28
1.50	31

Sambandet mellan mängden kolmonoxid, och tiden sedan den senaste cigaretten röktes kan sammanfattas med en regressionsmodell.

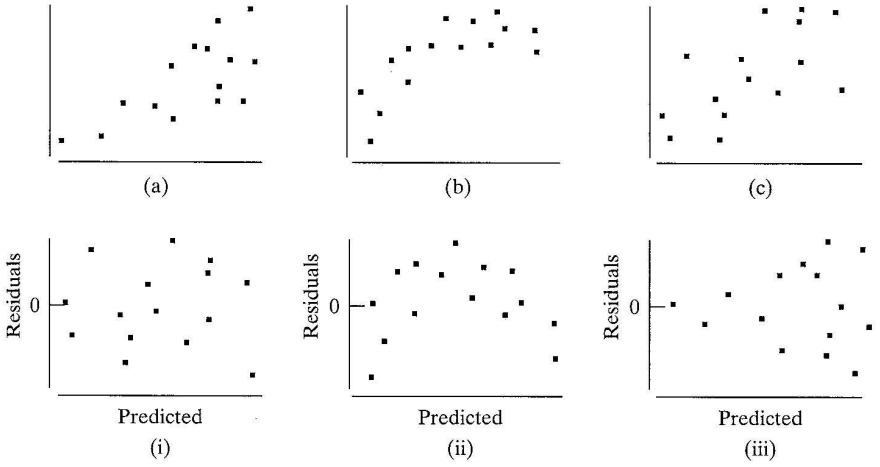
```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-14.194  -4.444  -1.698   5.566  16.696

Coefficients:
            Estimate      Std. Error
(Intercept)   45.205           4.393
 x            -6.007           1.844

---

Analysis of Variance Table

Response: y
          Df    Sum Sq
x          1    929.39
Residuals 10    875.53
```

(a) Det verkar rimligt att tro från början att vi ska ha ett avtagande samband mellan variablerna y och x. Gör ett lämpligt test eller konfidensintervall för att undersöka detta. Nivå 0.01. (2p)

(b) Ange förklaringsgraden i modellen. (1p)

(c) När man anpassar en regressionsmodell tittar man ofta på residualerna för att avgöra om modellen är lämplig. Residualerna är avvikelserna från varje observerad punkt till den skattade regressionslinjen. Nedan finns plottar på data (a), (b) och (c) där man anpassat en regressionsmodell samt plottar på residualerna från varje regressionsmodell.

Vilken skatterplot hör till vilken residualplot? (1p)



6 (a) I kvällstidningarna utlovas en viktminskning efter två veckor om man följer en ny diet. Data nedan beskriver vikten för sju personer före och efter två veckor där de följt dieten.

vikt före diet (kg)	vikt efter diet (kg)
99.2	94.1
83.7	82.7
91.1	92.1
60.1	63.4
80.8	78.0
71.5	74.4
78.7	79.8

Verkar dieten ha den påstådda effekten? Undersök med lämpligt test på nivå 0.05 eller konfidensintervall med motsvarande konfidensgrad. Glöm inte att motivera ditt val. Lämpligt normalfördelningsantagande får göras. (2p)

(b) I vissa situationer använder man F-test för att undersöka hypoteser. Beskriv en sådan situation och förklara varför F-fördelningen uppkommer. (2p)

Bioinformatik

7 Sekvensbioinformatik

Gör en global sekvensjämförelse (global alignment) mellan aminosyrasekvenserna RNA och NAR mer hjälp av Needlemann-Wunsch-algoritmen. Använd en linjär straff-funktion med konstant 8 och substitutionsmatrisen PAM120 där bland andra följande värden finns tabulerade.

	A	R	N
A	3	−3	−1
R	−3	6	−1
N	−1	−1	4

Vilken är den optimala uppställningen och vilken poäng har den? (4p)

8 Strukturbioinformatik

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

(b) Describe the *family* and *superfamily* levels in the Structural Classification of Proteins (SCOP). (2p)

Lycka till!