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

Tid: Torsdagen den 22 december, 2006 kl 14.00 - 18.00 i M-huset.
Examinator: Olle Nerman, tel 7723565.
Jour: Alexandra Jauhiainen, tel 7725380, 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 Kumaraswamy-fördelningen, ursprungligen skapad av den indiska hydrologen Poond Kumaraswamy, har följande täthetsfunktion,

$$f_X(x) = abx^{a-1}(1-x^a)^{b-1}\,, \quad 0 \leq x \leq 1.$$

Låt X vara Kumaraswamy-fördelad med $a = 2$ och $b = 2$.

(a) Beräkna E[X]. (1p)

(b) Beräkna Var[X]. (2p)

(c) Med *moden* för en kontinuerlig stokastisk variabel menas det x där täthetsfunktionen $f(x)$ har det största värdet. Beräkna moden för en Kumaraswamy-fördelad stokastisk variabel med parametrarna $a = 2$ och $b = 2$. (1p)

2 Antalet fel på en slumpmässigt vald sida i den 6:e upplagan av "Probability and Statitsics" av J. L. Devore antas vara Poisson-fördelat med väntevärde λ. Boken har totalt 796 sidor och felen på de olika sidorna antas vara oberoende. Låt X vara det totala antalet fel i boken.

(a) Om $\lambda = 2$, beräkna den approximativa sannolikheten att det totala antalet fel överstiger 1650. (2p)

(b) På förlaget där boken ges ut (Thomson) är man noga med kvalitén. Beräkna därför det största tillåtna värdet på λ (approximativt) om sannolikheten för att det ska finnas max 1000 fel ska vara mindre än 0.90. (2p)

3 Låt A vara en händelse som inträffar med sannolikheten 0.5 och låt X vara en stokastisk variabel. Den betingade fördelningen för X om A inträffar är Bin(10, 0.8) och den betingade fördelningen för X om A *inte* inträffar är Bin(10, 0.3).

(a) Beräkna sannolikheten för A *givet* $X = 8$. (2p)

(b) Beräkna E[X]. (2p)

Statistik

4 En industri destillerar luft som frusits till vätskeform för att producera syre, kväve och argon. Renheten hos det producerade syret tros vara linjärt avtagande av mängden orenheter i luften (föroreningstalet i ppm). Följande data insamlades under ett testförsök.

föroreningstal i ppm (x)	renhet % (y)
1.10	93.30
1.45	92.00
1.36	92.40
1.59	91.70
1.08	94.00
0.75	94.60
1.20	93.60
0.99	93.10
0.83	93.20
1.22	92.90
1.47	92.20
1.81	91.30
2.03	90.10
1.75	91.60
1.68	91.90

Vi tänker oss att y_j är en observation av en stokastisk variabel Y_j där

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j$$

En regressionanalys gjordes på materialet med följande resultat:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.846768 -0.195108  0.009844  0.270579  0.678471

Coefficients:
            Estimate Std. Error
(Intercept)  96.4546     0.4282
x            -2.9010     0.3056
---

Residual standard error: 0.4277 on 13 degrees of freedom
Multiple R-Squared: 0.8739, Adjusted R-squared: 0.8642


Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq
x          1 16.4908 16.4908
Residuals 13  2.3786  0.1830
```

(a) Undersök på nivån 0.01 hur föroreningstalet påverkar renheten hos syret. (2p)

(b) Vad betyder resultatet i (a)? (1p)

(c) Ett företag har jämfört efterfrågan på sitt nya hälsopreparat i tolv olika försäljningsdistrikt. I fem av distrikten såldes preparatet endast i hälsokostbutiker, medan det i de övriga distrikten även såldes i vanliga mataffärer. Försäljningsvolymen per invånare (y) registrerades och samtidigt togs tre förklarande variabler fram.

y	x_1	x_2	x_3
167	1	42.2	31.9
185	1	48.6	33.2
170	1	42.6	28.7
152	1	39.0	26.1
150	1	34.7	30.1
192	0	44.5	28.5
183	0	39.1	24.3
180	0	40.1	28.6
191	0	45.9	20.4
171	0	36.2	24.1
168	0	39.3	30.0
189	0	46.1	34.3

Här är x_1 är en indikator för om distributionen skedde endast via hälskokostbutiker, x_2 är urbaniseringsgraden och x_3 är relativ inkomst. Nedan följer analyser för varje föklaringsvariabel mot y. Vilken förklarande variabel skulle du välja om du endast fick ta med en? (1p)

Modell med x_1:

```
Call:
lm(formula = y ~ x1)

Residuals:
   Min     1Q Median     3Q    Max
-14.80 -11.45   1.60   7.50  20.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  182.000      4.432  41.068 1.76e-12 ***
x1           -17.200      6.866  -2.505   0.0312 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.73 on 10 degrees of freedom
Multiple R-Squared: 0.3856,     Adjusted R-squared: 0.3242
F-statistic: 6.276 on 1 and 10 DF,  p-value: 0.03116
```

Modell med x_2:

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5630  -7.5984   0.7488   8.8765  14.1886

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.7151    30.4371   2.356  0.04021 *
x2            2.4833     0.7296   3.404  0.00673 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.18 on 10 degrees of freedom
Multiple R-Squared: 0.5367,     Adjusted R-squared: 0.4904
F-statistic: 11.59 on 1 and 10 DF,  p-value: 0.006728
```

Modell med x_3:

```
Call:
lm(formula = y ~ x3)

Residuals:
     Min       1Q   Median       3Q      Max
-24.4382  -6.6035   0.2344  12.0392  17.2005

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 181.2347    31.8156   5.696 0.000199 ***
x3           -0.2258     1.1119  -0.203 0.843150
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.93 on 10 degrees of freedom
Multiple R-Squared: 0.004107,   Adjusted R-squared: -0.09548
F-statistic: 0.04124 on 1 and 10 DF,  p-value: 0.8432
```

5 I ett flertal forskningsartiklar inom det medicinska området har det rapporterats att patienter med kardiovaskulära sjukdomar eller cancer har hög närvaro av endotelceller eller rester därav i blodet. En metod som bland annat innefattar centrifugering följt av immunofluorescens-infärgning är användbar för att detektera endotelceller i blodet.

Venöst blod har provtagits från 7 patienter i en kontrollgrupp och mängden endotelceller (mg/ml) bestämts med metoden ovan. På samma sätt har blod från 18 patienter med ischemisk hjärtsjukdom (IHD) provtagits och mängden endotelceller bestämts.

	medelvärde	stickprovsstandardav.
Kontroll	1.447	0.347
IHD	3.355	0.979

Antag att data följer en normalfördelning.

(a) Undersök om varianserna kan anses vara lika i kontrollgruppen och gruppen med IHD-patienter. Nivå 0.1 (2p)

(b) Baserat på din slutsats i (a), undersök med lämplig metod om IHD-gruppen kan anses ha högre innehåll av endotelceller i sitt blod än kontrollgruppen. Nivå 0.01. (2p)

6 I skidskytte ska man i varje skjutning träffa med fem skott. Tänk att man följer en skidskytt under 250 skjutningar både på träning och tävling och registrerar antalet träffar i varje skjutning. Man arbetar enligt modellen att skidskytten träffar varje skott med sannolikheten p oberoende av andra skott. I en skjutning blir alltså X=antalet träffade skott binomialfördelad med parametrarna $n = 5$ och p.

(a) Härled maximum-likelihoodskattningen av p. Beräkna ett värde för $\hat{p}$ från observationerna nedan. (2p)

(b) Vi har fått följande observationer från de 250 skjutningarna:

Träffar	0	1	2	3	4	5
Antal obs.	25	65	33	21	60	46

Undersök med ett test på nivån 0.05 om vårt antagande om binomialfördelning är korrekt. Tips: Använd uttrycket för skattningen från uppgift (a) på den okända parametern p (om du gör det, hur blir då frihetsgraderna?). (2p)

Bioinformatik

7 Sekvensbioinformatik

(a) För att utföra parvis sekvensjämförelse används en substitutionsmatris, till exempel en PAM120-matris. Vilka egenskaper är önskvärda hos en substitutionsmatris? (1p)

(b) Vad är anledningen till att man använder olika substitutionsmatriser, exempelvis PAM40 och PAM120, vid parvis sekvensjämförelse? (1p)

(c) Kan man generalisera Needleman-Wunsch algoritm för global parvis sekvensjämförelse till global multipel sekvensjämförelse? Vilka praktiska begränsningar finns isåfall? (2 p)

8 Strukturbioinformatik

(a) What is the purpose of the DSSP program? Describe how main chain hydrogen bonds are calculated by DSSP. (2p)

(b) In protein modelling, what is a side chain rotamer? How are side chain rotamers used in protein modelling? (2p)

God Jul och lycka till!