Tentamen TMS145
Grundkurs i matematisk statistik och bioinformatik, 7,5 hp,
2011-04-26 kl 14.00-18.00.

Examinator: Olle Nerman, tel 772 35 65.
Jour: Malin Östensson, tel 0708347877

Hjälpmedel: valfri miniräknare, egen handskriven formelsamling (fyra A4 sidor) samt med skrivningen utdelade tabellsidor.

Maxpoäng: 32. För godkänt krävs minst 15 poäng totalt och minst 4 poäng på sannolikhetsteori- och statistikdelen vardera samt minst 3 poaäng på bioinformatikdelen.

Sannolikhetsteori

1. Vi har två händelser A och B med

$$P(A) = 0.1 \qquad P(B) = 0.4.$$

Sannolikheten att minst en av dem inträffar är 0.42.

(a) Vad är sannolikheten att både A och B inträffar?

1 p

(b) Vad är $P(B|A)$?

1 p

(c) Är A och B oberoende?

1 p

(d) Vad är sannolikheten att A inträffar men inte B?

1 p

2. Ur en välblandad kortlek dras sju kort, låt den stokastiska variabeln

$$X = \text{antalet klädda kort}$$

bland dessa sju, där klädda kort innefattar knekt, dam,kung och ess.

(a) Beskriv vilken sannolikhetsfördelning X har, samt ange vilka värden den kan anta och vilken sannolikhetssfunktion den har.

3 p

(b) Vad är sannolikheten att inte få något klätt kort?

1 p

3. Till ett apotek anländer kunder till två olika köer (recept och receptfritt) enligt två *oberoende* poissonprocesser med intensiteter 5 kunder/timma respektive 10 kunder/timma, dvs tiden från en kund till nästa är Exponentialfördelad med samma intensitet. Låt N_1 och N_2 beteckna antalet kunder under den första timmen, och låt Y_1 och Y_2 beteckna tiden tills den första kunden anländer (till resp. kö) under denna timma.

   (a) Beräkna sannolikheten att personalen får vänta minst 10 min på den första kunden ($P(\min(Y_1, Y_2) > 10)$).

   1 p

   (b) Vad är sannolikheten att det anländer minst 2 kunder till receptfritt-kön och inte någon till den andra kön under denna timma?

   1 p

   (c) Vad blir det förväntade antalet kunder under denna timma? ($\mathbf{E}[N_1 + N_2]$)

   2 p

Statistik

4. Låt $x_1, \ldots, x_{20}$ vara observationer från oberoende stokastiska variabler med täthetsfördelning

$$f(x; \alpha) = Cx^{\alpha-1}, 0 < x < 1,$$

för parametern $\alpha > 0$, C är en normaliseringskonstant.

   (a) Bestäm konstanten C så att f blir en giltig sannolikhetsfunktion.

   1 p

   (b) Beräkna väntevärdet för denna fördelning.

   1 p

   (c) Härled maximum-likelihood(ML) skattaren för α.

   2 p

5. En forskare vill skatta den teoretiska frekvensen av en viss biverkning av ett nytt läkemedel. Bland 1000 individer som prövar preparatet är det 19 personer som fått denna biverkning.

   (a) Vad är fördelen med intervall skattning istället för punktskattning av frekvensen?

   1 p

   (b) Ta fram en punktskattning av den teoretiska frekvensen (p) av individer i populationen som skulle få denna biverkning av detta läkemedel. Beräkna också ett nedåt begränsat 95%-igt konfidensintervall för p. Använd lämplig approximation.

   2 p

   (c) Testa nollhypotesen $H_0 : p = 0.03$ mot $H_1 : p < 0.03$ på nivån 5%. Vad blir resultatet?

   1 p

6. En gen med alleler a och A misstänks vara associerad med människors längd, för att undersöka detta tas stickprov ur de tre genotypkategorierna:

	AA	Aa	aa
n	10	20	10
$\overline{y}$	183.01	173.94	171.53
s^2	79.00	99.89	90.27

I tabellen betecknas stickprovsstorleken i de tre grupperna av n, $\overline{y}$ är stickprovsmedelvärdet av personernas längd i respektive grupp, och s^2 betecknar stickprovsvariansen av längden y. Låt

$$x = \begin{cases} 0, & \text{om aa} \\ 1, & \text{om Aa} \\ 2, & \text{om AA} \end{cases}$$

vara förklaringsvariabel. Utför en regressionsanalys med

$$S_{xx} = 20, S_{xy} = 114.85, \overline{y} = 175.61$$

$$SS_T = 4192.04, SS_R = 659.5,$$

(a) Ge ett uttryck för den linjära regressionsmodellen med längd (y) som svarsvariabel, baserat på informationen ovan. Tolka koefficienterna i termer av problemets frågeställning.

2 p

(b) Avgör om det finns ett signifikant linjärt samband mellan antalet A-alleler och en persons längd. Motivera ditt svar.

2 p

Bioinformatik

7. Sequence Alignment

   (a) i. What is the Hamming distance between strings "RNA" and "NAR"? Explain your answer.

   ii. What is the Levenshtein distance between strings "RNA" and "NAR"? Explain your answer.

   (b) i. 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 "RNA" and "NAR".

   ii. 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 "RNA" and "NAR".

4 p

8. Structural Bioinformatics

   (a) i. Draw a sketch that shows the phi torsion angle in a protein's main chain. Identify all atoms in your sketch, and use subscripts to identify the residue to which each atom belongs.

   ii. Explain why the allowed phi angles for a glycine residue are different to the allowed phi angles for other amino acid residues.

2 p

(b) Describe the planarity checks performed by PROCHECK.

2 p

Lösningsförslag TMS145 26 April 2011

1 a $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$0.42 = 0.1 + 0.4 - P(A \cap B)$

$P(A \cap B) = 0.5 - 0.42 = 0.08$

b $P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.08}{0.10} = 0.8$

c A och B oberoende om

$P(A \cap B) = P(A)P(B)$

$0.08 \neq 0.1 \cdot 0.4$

$\therefore$ A och B är INTE oberoende!

d $P(A \cap B^c) = P(A) - P(A \cap B)$ (alt. $P(A \cup B) - P(B)$)

$= 0.1 - 0.08 = 0.02$

---

2 a • $X \sim \text{HypGeom}(n=7, M=16, N=52)$

• X kan anta värdena $0, 1, \dots, 7$

• $P(X=x) = \frac{\binom{12}{x}\binom{40}{7-x}}{\binom{52}{7}}$

b $P(X=0) = \frac{\binom{16}{0}\binom{36}{7}}{\binom{52}{7}} = \frac{1 \cdot 8347680}{133784560} = 0.062$

(Alt. {inga kläckta kort} är gynnsamma utfall:

$\binom{36}{7}$

Totalt antal möjliga utfall: $\binom{52}{7}$

$P(\text{inga kläckta}) = \frac{\#\{\text{gynnsamma}\}}{\#\{\text{möjliga utfall}\}} = \binom{36}{7} / \binom{52}{7}$

3 $N_1 \sim Poi(\lambda_1 = 5 \text{ kunder}/h)$ $Y_1 \sim Exp(\lambda_1)$

$N_2 \sim Poi(\lambda_2 = 10 \text{ kunder}/h)$ $Y_2 \sim Exp(\lambda_2)$

a $P(\min(Y_1, Y_2) > 10)$

$= P(Y_1 > 10, Y_2 > 10) = [\text{oberoende}]$

$= P(Y_1 > 10) P(Y_2 > 10) = [\text{Exp. fördelade}]$

$= (1 - F_{Y_1}(10))(1 - F_{Y_2}(10))$

$= e^{-\lambda_1 \cdot 10/60} \cdot e^{-\lambda_2 \cdot 10/60} = e^{-(\lambda_1 + \lambda_2) 10/60}$

$= e^{-150/60} = \underline{0.082}$

b $P(N_2 \geq 2, N_1 = 0) = P(N_2 \geq 2) P(N_1 = 0)$

$P(N_2 \geq 2) = 1 - P(N_2 \leq 1) = 1 - \sum_{x=0}^{1} \frac{e^{-\lambda} \lambda^x}{x!}$

$= 1 - \left(\frac{e^{-10} \cdot 10^0}{0!} + \frac{e^{-10} 10^1}{1!} \right) = 1 - e^{-10}(1+10)$

$= 0.9995$

$P(N_1 = 0) = \frac{e^{-5} \cdot 5^0}{0!} = e^{-5} = 6.7379 \cdot 10^{-3}$

$= 0.9995 \cdot 6.738 \cdot 10^{-3} = \underline{6.735 \cdot 10^{-3}}$

c Eftersom N_1 och N_2 är oberoende blir

$E[N_1 + N_2] = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} (n_1 + n_2) \frac{e^{-10} 10^{n_2}}{n_2!} \frac{e^{-5} 5^{n_1}}{n_1!}$

$= \sum_{n_2} \frac{e^{-10} 10^{n_2}}{n_2!} \underbrace{\left(\sum_{n_1} n_1 \frac{e^{-5} 5^{n_1}}{n_1} \right)}_{E[n_1] = 5} + \sum_{n_1} \frac{e^{-5} 5^{n_1}}{n_1!} \underbrace{\left(\sum_{n_2} n_1 \frac{e^{-10} 10^{n_2}}{n_2} \right)}_{E[n_2] = 10}$

$= 5 \underbrace{\sum_{n_2} \frac{e^{-10} 10^{n_2}}{n_2!}}_{=1} + 10 \underbrace{\sum_{n_1} \frac{e^{-5} 5^{n_1}}{n_1!}}_{=1} = \underline{15}$

4. $f(x, \alpha) = C x^{\alpha-1}$, $0 < x < 1$

a $\int_0^1 C x^{\alpha-1} dx = C\left[\frac{x^\alpha}{\alpha}\right]_0^1 = \frac{C}{\alpha}(1-0)$

$\int_0^1 f(x) dx = 1 \iff \underline{C = 1/\alpha}$

b $E[X] = \int_0^1 x \cdot f(x) = \int_0^1 x \cdot \alpha x^{\alpha-1} dx$

$= \int_0^1 \alpha x^\alpha dx = \alpha\left[\frac{x^{\alpha+1}}{\alpha+1}\right]_0^1 = \underline{\frac{\alpha}{\alpha+1}}$

c Likelihood funktionen:

$L(\alpha; x_1, \ldots, x_{20}) = \prod_{i=1}^{20}(\alpha x_i^{\alpha-1}) = \alpha^{20}\left(\prod_{i=1}^{20} x_i\right)^{\alpha-1}$

$\ell(\alpha; x_1, \ldots, x_{20}) = 20 \ln(\alpha) + (\alpha-1)\sum_{i=1}^{20} \ln(x_i)$

$\frac{\partial \ell}{\partial \alpha} = \frac{20}{\alpha} + \sum \ln(x_i) = \frac{20}{\alpha} + \ln\left(\prod x_i\right)$

$\hat{\alpha}_{ML} = -\frac{20}{\ln(\prod x_i)} = \frac{20}{\ln(\prod 1/x_i)} = -\frac{20}{-\sum \ln x_i}$

5

b $\hat{p} = \frac{x}{n} = 19/1000$

$X \sim Bin(1000, p)$

p^* skattas m. $\hat{p} = x/n$

$p^* = \frac{X}{n} \sim$ appr $N\left(p, \frac{p(1-p)}{n}\right)$ Var (p^*) sk.m. $\frac{\hat{p}(1-\hat{p})}{n}$

Konfidensintervall

$$\hat{p} + a \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{19}{1000} - a\sqrt{\frac{19/1000\,(1+19/1000)}{1000}} = 0.0119$$

$a = 1.64 \quad (\alpha = 0.05,\ z_{0.95} = 1.64)$

Dvs intervallet blir

$(0,0119\ ;\ 1)$

c $z = \frac{p^* - p_0}{\sqrt{\frac{p^*(1-p^*)}{n}}} \Rightarrow z = \frac{0.019 - 0.03}{0.004317} = -2.547$

Ensidigt test på nivån $\alpha = 0.05 \Rightarrow$ Förkasta H_0 om $z < -1.64$, alltså kan H_0 förkastas.

---

6 a $y_i = \alpha + \beta x_i + \varepsilon_i \quad , \varepsilon_i \sim N(0, \sigma^2)$

α, β skattas med minsta kvadrat metoden:

$\hat{\beta} = S_{xy}/S_{xx} = 114.85/20 = \underline{5.7425}$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 175.61 - 5.7425 \cdot 1 = \underline{169.8675}$

b) F-test:

$V = \frac{SS_R/1}{SS_E/(140-2)} = \frac{659.5/1}{(4192.04 - 659.5)/38} = 7.09$

alternativt: P value: 0.0112

T-test:

$t = \frac{\hat{\beta} - \beta_0}{s/\sqrt{S_{xx}}} = \frac{5.74}{\sqrt{\frac{SS_E}{38}}/\sqrt{S_{xx}}} = 2.66$ P-val = 0.0056