Jan., 08, 03

Homework Set No. 2

Answer the following questions. Please provide a type in copy of the solution.

Q#1: A protein database contains 10 protein sequences. 4 proteins in this database are homologues and form a protein family:

a) How many different possible ways are there of forming this family.

b) Three other proteins also constitute a family. How many different ways are there of placing the 10 proteins into the two families.

Q#2: The PFAM profile database contains 2700 domain families:

a) How many different four domain protein are possible.

b) How many four domain proteins are possible in which no two domains in the proteins are homologous to each other.

Q#3: Consider only standard nucleotides (A,T,G, or C) and standard amino acids (20 of them)

a) How many unique 30-base long DNA sequences are there?

b) How many unique 10-residue long peptide sequences are there?

c) Considering that the correspond DNA sequence of a 10-residue long peptide is 30 nucleotide long, should the two numbers obtained in a.) and b.) be the same? If not, why?

Q#4: How often an TaqI site would be expected to appear by chance in a random sequence.

Q# 5: A base calling procedure is very acurate in determining the nucleotides in a DNA sequence, in that it correctly identifies each base with high probability and only rarely misclassifies bases. Let E_A,E_C, E_G and E_T be the events that base under study is identified as an A,C,G and T respectively, and let F_A, F_C,F_G and F_T be the events that base under study is actually an A,C, G and T respectively. Suppose that prior to any analysis being carried out it is assumed that:

$$\begin{array}{lllll} p(F_A) & = & p_A & = & 0.30 \\ p(F_G) & = & p_G & = & 0.20 \end{array} \qquad \begin{array}{lllll} p(F_C) & = & p_C & = & 0.20 \\ p(F_T) & = & p_T & = & 0.30 \end{array}$$

and the conditional probabilities of (mis) classification of a base, give that its actual type are given by the following tables

		Is Called As			
		A	C	G	T
The Base	A	0.900	0.025	0.025	0.050
	C	0.025	0.850	0.100	0.025
	G	0.025	0.100	0.850	0.025
	T	0.050	0.025	0.025	0.900

so that, for example, from the top row

$$p(E_A|F_A) = 0.900 \qquad p(E_C|F_A) = p(E_G|F_A) = 0.025 \qquad p(E_T|F_A) = 0.05$$

and so on.

i). Using the Total probability formula, compute the probability that an unknown base under analysis is classified as $i \in \{A, C, G, T\}$, that is, compute

$$p(E_i) = \sum_{j \in \{A,C,G,T\}} p(E_i|F_j)p(F_j) \qquad \text{for each } i \in \{A, C, G, T\}$$

ii). Compute, using Bayes Theorem or the conditional probability formula, the conditional probability that a base is actually an A, given that is classified as an A i.e., $p(F_A|E_A)$.
Compute also the three conditional probabilities that a base is actually an A, given that it classified as C.

Q#6: Count data from two DNA sequences was collected

	Nucleotide				
Sequence	A	C	G	T	Total
1	250	140	180	230	800
2	320	270	310	300	1200
Total	570	410	490	530	2000

A test of the null hypothesis H_0, that the marginal probabilities of the four nucleotides are identical for both sequences, is required.

i). Complete the table of expected counts

$$e_{ij} = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, \ldots, r, \ j = 1, \ldots, c$$

Assuming H_0 is true, where $n_{i.}$ is the total of the i^{th} row, $n_{.j}$ is the total of the j^{th} column, and n is the total number of observations.

ii). Compute the Chi-squared statistic χ^2.
Recall that, here, the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$

iii). Carry out a test of H_0 at the significance level of $\alpha = 0.01$.
Note: You could verify your result by using Minitab or R. The R-code to solve this problem is:
$x < -c(250, 140, 180, 230, 320, 270, 310, 300)$
$data.matrix < -matrix(x, ncol = 4, byrow = TRUE)$
$c < -colSums(data.matrix)$
$r < -rowSums(data.matrix)$
$gt < -sum(data.matrix)$
$p < -matrix(0, ncol = 4, nrow = 2)$
$for(i\ in\ 1:2)$
$\{for(j\ in\ 1:4)$
$\{p[i,j] < -c[j] * r[i]/gt^2\}\}$
$chi.test < -chisq.test(data.matrix, p)$

Here some information about the chi-square functions available in R:
$dchisq(x, df, ncp = 0, log = FALSE)$
$pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)$
$qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)$
$rchisq(n, df, ncp = 0)$
Arguments:
x, q : vector of quantiles.
p: vector of probabilities.
n: number of observations. If '$length(n) > 1'$, the length is taken to be the number required.
df: degrees of freedom.
ncp: non-centrality parameter. For 'rnchisq', '$ncp = 0'$ is the only possible value.
log, log .p: logical; if TRUE, probabilities p are given as $log(p)$.
lower.tail: logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.