

CBB 200 Problem set 1

Due Mar 20

In the following problem set solve:

2 from category *A*

2 from category *B*

3 from category *C*

1 from category *D*

Pr. A.1 Given n i.i.d. random variables that are normally distributed with mean μ and variance 1 and the null hypothesis $\mu = 0$ and the alternative $\mu > 0$, find the p-value associated with an observed value of 4.4 with $n = 10^2, 10^4, 10^5, 10^6$.

Pr. A.2 Consider duodenal cancer patients and normal controls in the following table

Phenotype	Ulcer patient	Normal
A	186	279
B	38	69
AB	13	17
O	284	315

Let p, q, r be the vector of allele frequencies among patients, controls, and combined. Test the hypothesis $p = q$. Test the Hardy-Weinberg equilibrium null hypothesis for the cancer and normal patients.

Pr. A.3 Given a multinomial model with n trials and m categories with a probability p_i for each category i the number of categories having d or more observations is a well studied statistic W_d . This statistic has mean

$$\lambda = \sum_{i=1}^m \left[\sum_{k=d}^n \binom{n}{k} p_i^k (1 - p_i)^{n-k} \right].$$

The quantities in the inner sum is the probability that category i has more than d counts. If the variance of W_d is close to λ it is approximately Poisson with parameter λ .

A study tabulated 66 mutations in 141 amino acids of the hemoglobin α -chain. Of these 141 amino acids 16 show two or more mutations.

Assume $p_i = 1/141$. Compute the p-value for W_2 under the Poisson approximation and exactly.

- B.1** a) Consider a Markov chain with state space $\{1, 2, 3\}$ and transition matrix

$$\begin{bmatrix} .4 & .2 & .4 \\ .6 & 0 & .4 \\ .2 & .5 & .3 \end{bmatrix}.$$

What is the probability in the long run that the chain goes to states 1, 2, 3. Compute this using both the power and the eigenvector methods.

- b) Do the same for

$$\begin{bmatrix} .2 & .4 & .4 \\ .1 & .5 & .4 \\ .6 & .3 & .1 \end{bmatrix}.$$

- B.2** Given the Markov chain with state space $0, \dots, 5$ and transition matrix

$$\begin{bmatrix} .5 & .5 & 0 & 0 & 0 & 0 \\ .3 & .7 & 0 & 0 & 0 & 0 \\ 0 & 0 & .1 & 0 & .9 & 0 \\ .25 & .25 & 0 & 0 & .25 & .25 \\ 0 & 0 & .7 & 0 & .3 & 0 \\ 0 & .2 & 0 & .2 & .2 & .4 \end{bmatrix}.$$

Draw the corresponding graph. Is the chain irreducible. What is the steady-state transition matrix. Break the chain up and compute steady-state probability for states within the broken subchain.

- Pr. C.1** Consider the simple random walk model in class (probability of going up 1 as p and probability of going down 1 as q) with $h = 0$, $b = 1$, and $a = -L$ (where L is positive). Use

$$w_h = \frac{e^{\theta^* h} - e^{\theta^* a}}{e^{\theta^* b} - e^{\theta^* a}}$$

to compute the probability the walk eventually reaches 1 rather than $-L$. Show for $p < q$ the limit of this probability as $L \rightarrow \infty$ is $\frac{p}{q}$

Pr. C.2 Given the solution to the above problem show that this implies the probability that a walk reaches y is simply $\binom{p}{q}^y$ for any positive integer y .

Pr. C.3 Consider the following simple symmetric Markov matrix (it is an example of a PAM of mutation probability matrix) the matrix is of size 20 and $M_{jj} = .99$ and $M_{kj} = .01/19$ for $j \neq k$. Assume the initial probabilities of the amino acids are uniform $p_j = .05$ Run the matrix over $n = 260$ time step, this results in a matrix M^n (this is our units of evolution). We now define a new matrix of substitution probabilities as

$$q(j, k) = p_j M_{jk}^n.$$

We also define a scoring matrix where $S_{jj} = 12$ and $S_{ij} = -1$ for $i \neq j$. Use this matrix and BLAST theory to compute the mean step size and expected number of high scoring excursions.

Pr. C.4 Give a detailed explanation of the the complete BLAST output in example 9.5.1 in Statistical Methods for Bioinformatics by Ewans and Grant or run your own BLAST search and give a detailed explanation.

D.1 There are mn patients in a hospital ward and their beds are arranged as n rows by m beds (m, n are even). Suppose one of the patients in bed $m/2$ in row $n/2$ becomes infected and can infect a patient in one of the four neighboring beds. Once a patient is infected they stay contagious for k days and then recover and are cured. For every day of infection the probability that each neighbor gets infected is τ . For any day t what is on average $I(t)$ – number of infected patients, $S(t)$ – the number of patients never infected, and $R(t) = 1 - I(t) - S(t)$ – the number of patients that recovered.

Plot this model for $m = n = 100$ and $\tau = .2$ and $m = n = 10$ and $\tau = .2$.

Pr D.2 Goto the following webpage

<http://www.stat.duke.edu/sayan/webPub/gsea.pdf>

Read the paper on gene set enrichment analysis. Re-implement the algorithm with a weighted Man-Whitney statistic as the relevant computation on the random walk and Pearson correlation as the correlation statistic. Run this on simulated data.