

GENOME 560, Spring 2012

Problem Set #3

(Due May 24th 9:00am)

1. [50 points] Maximum Likelihood Estimation

In lecture last Thursday, we calculated the maximum likelihood estimate (MLE) of the proportion of recombinant gametes between two heterozygous loci. Now, we are interested in applying this method to obtaining the MLE of the number of spontaneous deleterious mutations in humans. Say that you developed a potential Nobel prize worthy method of counting spontaneous deleterious mutations. You applied it to ten individuals and obtained the following data:

Individual	Number of deleterious mutations
1	4
2	3
3	0
4	1
5	9
6	2
7	3
8	8
9	0
10	6

- Write out the likelihood function for this data. (Hint: we discussed in class the probability distribution that this type of data follows.)
- Calculate the MLE of the average number of spontaneous deleterious mutations per individual from the data above. You can do this in any of the three ways that we discussed in class (graphically, numerically or calculus).
- Plot the log-likelihood of the data

2. [50 points] The t-test

You are interested in the transcriptional changes during early stages of the innate immune response. You obtain lymphoblast cell lines from 10 individuals and for each

one measure expression levels at baseline (untreated) and following treatment with the drug immiquimod (which is a TLR8 agonist). The following table shows gene expression levels for a particular transcript.

Individual	Baseline	Stimulated
1	-0.24	1.74
2	0.25	2.1
3	1.12	1.65
4	-0.06	2.65
5	0.46	3.11
6	0.17	2.31
7	0.02	1.87
8	1.10	3.21
9	0.55	2.19
10	0.98	1.75

- (a) Perform a one sample t-test to test the hypothesis that baseline expression levels are significantly different than zero. Clearly state the null and alternative hypotheses and submit R code, test statistic value and p-value.
- (b) Use a paired t-test to test the hypothesis that gene expression levels are significantly different between baseline and stimulated conditions. Again, clearly state the null and alternative hypotheses and submit R code, test statistic value and p-value.
- (c) An alternative way of analyzing the data as opposed to a paired two sample t-test (part b) is to create a new phenotype for each individual defined as the difference between stimulated and baseline expression. Formally, let x_i and y_i denote the expression level for the i -th individual in baseline and stimulated conditions, respectively. Then define $z_i = y_i - x_i$. Perform a one sample t-test on the vector of z_i values. Clearly state the null and alternative hypotheses and submit R code, test statistic value, and p-value. How does your result compare to that obtained from part b?