Homework 0B - R Genetics Sample Session

CS/HG 124/224

Due: April 15th, 2014

All assignments must be submitted by a hard copy to TA's office by 4 pm on the day they are due.

1 Problem 1

Imagine you toss a fair coin 20 times. Let A be the event that at least one coin toss comes up heads. What is the probability that A occurs? (Hint: Think about the complement of A, A^c.) Lets do this calculation using R. Write a R code to calculate the probability.

2 Problem 2

It is known any one screw produced by a certain company will be defective with a probability of 0.01. The company sells the screws in packages of 10 and offers a money-back guarantee if at least 2 of 10 screws is defective. What proportion of packages sold must the company replace? (Hint: Again, think about the binomial distribution and let X be a random variable that counts the number of defective screws.) Lets do this calculation using R. Write a R code to calculate the probability.

3 Problem 3

In this problem, we want to discover if a SNP is potentail causal to certain gene expression. To examine the potential causal relationship between SNP S and expression level E, we assume the following linear relationship between the two: $E = \alpha S + \epsilon$. We use an arrow notation to signify potential causality ($\rightarrow$) and the negation ($\nrightarrow$) as no potential causality. Under the null hypothesis of no potential causal relationship between the SNP and expression levels ($S \nrightarrow E$), we expect $\alpha = 0$ (H_0). Under the alternate hypothesis of a potential causal relationship ($S \rightarrow E$), we expect $\alpha \neq 0$ (H_1). To decide between these hypotheses, we calculate the likelihood ratio statistic $x = -2 \log \frac{\mathcal{L}(H_0)}{\mathcal{L}(H_1)}$ (we decide to perform likelihood ratio test). If we assume that under the null hypothesis, the distribution of statistic follows χ^2 distribution with 1 degree of freedom and a non centrality parameter of 0, calculate the p-values of the following observed statistics of each candidate SNP. If we use the p-value threshold of 0.001, through our likelihood ratio test, which SNP can be claimed as causal SNP? Write R codes for each p-value calculation.

SNPs	LRT Statistics	p-value	Causal? (Yes/No)	R code for p-value calculation
SNP1	0.05			
SNP2	2.5			
SNP3	17.5			
SNP4	0.15			
SNP5	6.7			
SNP6	9.8			

Table 1: Likelihood Ratio Test Result Table

4 Problem 4

The following problem requires that you do it in R. This is actually a real biology question but we will give you everything you need so that an understanding of the biology is not necessary.

In 2002, Brem et. al. performed a linkage analysis on the gene expression of some 6000 genes in yeast and found 385 significant linkages between the expression of a gene and a region of the genome. One of the goals of the analysis is to see if certain parts of the genome are enriched in these 385 significant linkages. They called these regions "regulatory hotspots" and they might contain regulators that control the gene expression of a large number of genes. We will study how they identified these regions.

4.1 Part A

Imagine that we have divided the genome up into 611 bins of 20 kb each. Assuming that linkages are uniformly distributed across the genome (any given linkage has an

equal probability of being in one of the bins), what's the probability that a linkage occurs in a bin?

4.2 Part B

Let X be the random variable counting the number of significant linkages a bin would contain. What's the probability that a bin would contain $X > 1$ significant linkages. (Hint: Assume a binomial distribution with $n = 385$ and p equal to what was computed in Part A and use pbinom. If you couldn't compute Part A, use $p = 0.001$.)

4.3 Part C

Again, let X be the number of significant linkages a bin would contain. Compute and plot the probability density functions for $X = x$ and the cumulative distribution functions for $X > x$ where $x \in \{1, 2, 3, 4, 5\}$. Attach the plot to a hard copy.

4.4 Part D

We often model rare events using the Poisson distribution. In this case, we can model the rare event of a significant linkage occuring in a bin. At what rate do significant linkages fall into a bin? (Hint: Rate is the ratio between the number of significant linkages and the number of bins.)

4.5 Part E

The rate is the only parameter you need for the Poisson distribution. Compute and plot the probability density functions for $X = x$ and the cumulative distribution functions for the Poisson distribution for $X > x$ where $x \in \{1, 2, 3, 4, 5\}$. Does it look similar to Part B? Attach the plot to a hard copy.