COMP 364 - Tools for the Life Sciences

Assignment Solutions #4, 2016

M Hallett, M Ghadie

Question 1: [45% of assignment] Simple binomial approximations.

You are a high powered scientist that buys and sells thousands of graduate students daily. Your group generated a dataset of RNA-seq gene expression profiles for 900 breast carcinoma. You assign three students to compete on the same project to do the analysis. Only two will get a Ph.D. and will continue on in science. (If you could look into the future, you would see that the third will take a job at Google and end up making 18x as much as you do; eventually they will buy all the houses that surround your home and build a tar factory.)

So in this highly competitive scientific world, you decide to award a Ph.D. to the two doctoral candidates that you believe have the best results. Each of three candidates choses a different types of test (t-test, wilcoxon, Kolmogorov-Smirnoff, … it doesn't matter really what their choice are) to measure every one of the 20,000 probes for differentially expression between good and bad outcome. All of them use the same level of significance p-value < 0.001. Student A finds 100 probes differentially expression, Student B finds 250 and Student C finds 750. There are 10 genes in common between A and B, 20 between A and C, and 10 between B and C.

a. For each of the three intersections (A and B, A and C, B and C), show how you can ask the question "Is the observed intersection significant?" (10, 20, 10 respective) using the bag & ball analogy that I have been droning on about over the past week. Make sure that you label what all the parameters are and what they represent (white balls, red balls, trials, successes etc.). Here use the binomial distribution (see below) to compute these probabilities

b. Who would you fire and why? Justify your answer.

So here is the help returned from `? binom.test()`

```
binom.test(x, n, p = 0.5,
        alternative = c("two.sided", "less", "greater"),
        conf.level = 0.95)
```

Arguments

x	number of successes, or a vector of length 2 giving the numbers of successes and failures, respectively.
n	number of trials; ignored if x has length 2.
p	hypothesized probability of success.
alternative	indicates the alternative hypothesis and must be one of "two.sided", "greater" or "less". You can specify just the initial letter.
conf.level	confidence level for the returned confidence interval.

Question 2: [30% of assignment] But now with hypergeometric

Repeat Question #1 but this time you won't allow for replacement (once a ball is chosen, it is removed from the bag).

You can use the functions `phyper()(or dhyper())` in R. Repeat Questions 1b-d but using the hypergeometric distribution.

Question 3: [25% of assignment] MuTect

For the probabilistic model described in class, consider the equation for

$$Pr[\, b_i \mid e_i,\, r,\, m,\, f\,]$$

which was divided into three parts (b_i = r, b_i = m, otherwise).

Show that when you add up all of these three events, they equal 1 (as a probabilistic model should).