Four-choice Probability

Day 3 learning objectives covered:

- Compute summary statistics from a dataset
- Use computer simulations to build intuition about random variables
- Simulate random distributions
- Articulate and encode the Null Hypothesis
- Generate a simulated distribution under the Null Hypothesis
- Compare simulated results to data and obtain a p value
- Accept or reject a hypothesis based on the outcome of a simulation experiment

See Matlab code in ProbHist.m for a solution.

Step 1: Write a function for our "random number" game:

Have N people pick a number from 1 to R, inclusive. What is the probability, p, of K or more choices of any single number? Your function should take inputs R, N, K and a 4th variable, nSims (# of simulations to run) and return the probability, p.

p = MyFunction(R, N, K, nSims);

This example is from actual class data taken on 11 August 2014. The instructions were to pick at random a number from 1 to 4. Here is the actual data from 65 respondents:

5 people chose number 1
13 people chose number 2
35 people chose number 3
12 people chose number 4

Is this distribution random? How would you figure this out?

Step 2: Are students biased towards the number 3?

The hardest part of this exercise is defining what your Null Hypothesis is and what class of outcomes you would consider "weird." You need to have a precise definition so that you can count how often they occur when you generate a bunch of simulations under the Null Hypothesis (H_0). In this step, we choose H_0 to be “The number 3 is chosen as often as any other number.” The Alternative Hypothesis (H_A) (the hypothesis you are testing) is “The number 3 is chosen more often than any other number.”

Under this choice of H_0 and H_A, you would count any event where the number 3 is chosen 35 times or more as an event as or more extreme than the one you observed. On the other hand, an event where only 2 people choose the number 3 and 35 people choose the number 1 would be considered as being in line with H_0.

What is the probability of observing the above event (or one more extreme) given H_0? (i.e. What is your p-value?) What is your conclusion?

Step 3: Are students biased towards one particular number?

Consider a different choice of H_0 and H_A. We might not be interested in the number 3 especially, but only in the general event of one number being picked much more often than the others. In this case, what is H_0? What is H_A? What is your p-value?

Step 4: What result would convince you that a bias exists? In a class of 75 people, the smallest number of picks of any given number which would cause you to reject the null hypothesis (randomness) at a significance level of 0.05?

Step 5: Which formulation of H_0 and H_A would you have picked? Can you think of another way to phrase H_0 and H_A that would affect the p value? What are the advantages and disadvantages?

Originally part of Rattus binomialis, created by RTB
RTB separated this out as a separate module, 20 May 2014
MIS added Day 3 learning objectives and details about choosing H0 and HA, 31 Jul 2014
MIS changed numbers to results from August class poll, 13 Aug 2014