David Youngberg
BSAD 210—Montgomery College

LECTURE 36: DISCRETE PROBABILITY FUNCTIONS

I. Hypergeometric Distributions
   a. In a binomial distribution, we discussed the assumption of each trial as independent. For example, the sample taken is less than 5% of the population or there is replacement.
   b. What happens if your sample is more than 5% and there's no replacement? If each trial affects the likelihood of success, we need to use a different discrete probability function: hypergeometric.
   c. The mean is:

$$\mu = \frac{nR}{N}$$

      i. Where *N* is the population size
      ii. *R* is the number of successes in the population
      iii. *n* is the sample size.

II. Hypergeometric
   a. The command for hypergeometric is "=HYPGEOM.DIST" and it will tell you the chance something will happen given a particular set of values you put it. It requires four different pieces of information. In order they are:
      i. *Sample_s*: The number of successes in the sample (called x in the notes).
      ii. *Number_sample*: The size of the sample, or number of trials (called n in the notes).
      iii. *Population_s*: The number of successes in the population (called R in the notes).
      iv. *Number_population*: The size of the population (called N in the notes).
      v. *Cumulative*: type in either a 1 or a 0 for this value. (Or TRUE or FALSE.)
   b. Suppose you're buying 20 back-up generators from ValuCorp. ValuCorp generators are quite cheap but they are unreliable (and their return policy is a pain). You don't have time to test to make sure each one is working before you have to sign the contract for payment so you test 5 of them. If 3 generators are faulty, what is the chance that you will find 1 faulty generator? (This would justify you being able to take more time to test them all.)

i. First recognize that this is a hypergeometric distribution. For every generator tested, the probability of finding a faulty one (which is a "success" in this case) changes. Each probability is not independent.
ii. Type "=HYPGEOM.DIST(1,5,3,20,0)" and press ENTER. You should get about 0.4605, or 46.05% chance.
iii. Does this mean there's more than a 50% chance of not detecting any faulty generators? Note necessarily. Remember, there's a chance you can detect two faulty ones, or all three.
iv. Type "=HYPGEOM.DIST(0,5,3,20,0)" and press ENTER. You should get about 0.3391, or 33.91% chance.
v. That might be enough to satisfy you, it might not. Try testing 6, 7, and 8 generators instead. Note how the chance of finding zero faulty ones falls.

III. Poisson Distribution

a. This type of distribution describes the number of times some event occurs during a particular interval (such as time, distance, area, volume, etc). Unlike other distributions, there can be any number of occurrences (successes).
    i. Examples: number of returns in an hour; number of strawberries in a patch that don't pass quality control; number of lost golf balls per year at a mini-golf course.
b. Requirements
    i. Mean must be the same for each interval.
    ii. Intervals cannot overlap.
    iii. Occurrences in each interval must be independent.
c. If we know how often something occurs on average, we can use Poisson to figure out how often something other than the average occurs.
    i. Because the Poisson distribution begins with knowing the average number of events, there is no equation for the average number of events.
d. The standard deviation is:

$$\sigma = \sqrt{\lambda}$$

    i. Where λ is the average number of events that occur in the period in question.

IV. Poisson

e. The command for Poisson is "=POISSON.DIST" and it will tell you the chance something will happen given a particular set of values you put it. It requires three different pieces of information. In order they are:
    i. *x*: The number of successes in the interval (called x in the notes).

ii. *Mean*: The average number of events that occur in the interval (called λ in the topic notes).
iii. *Cumulative*: type in either a 1 or a 0 for this value. (Or TRUE or FALSE.)

f. During WWII, Germany launched a series of bombing raids on England, especially focusing on London. Between June 1944 and March 1945, 537 flying-bombs fell on South London. At this time, South London was divided into 576 regions of equal areas, meaning an average of 0.9323 bombs per region. The Nazis couldn't "aim" these bomb droppings; they were randomly dropped.[1]
    i. While we have a roughly one bomb a region, knowing how likely a region might get two or three bombs would help with emergency preparedness. How likely is it that a region will get two bombs?
    ii. First, recognize that this is Poisson: the regions don't overlap, a bomb landing in one region doesn't make it more or less likely another will fall in that region (independent), and the average is the same for each region.
    iii. Type "=POISSON.DIST(2,0.9323,0)" and press ENTER. You should get about 0.171, or 17.1% chance.
    iv. What is the chance that any particular region will suffer 3 or more bombs? You should get 0.0683, or about 6.83%.

[1] During the war, British statistician R.D. Clarke demonstrated that the bombings followed the Poisson distribution, thus suggesting that the bombs fell randomly and were not aimed.
https://data.princeton.edu/wws509/notes/c4s1