David Youngberg
BSAD 210—Montgomery College

LECTURE 24: DISCRETE PROBABILITY DISTRIBUTIONS

I. Discrete probability distributions
   a. A *distribution* lists all the possible results with the frequency of each result. It's typically presented in a graph form.
   b. A *discrete probability distribution* is distribution of data made up the results from a discrete random variable (which has outcomes of a small range of whole numbers).
   c. Some examples of a discrete random variable:
      i. The result of a flip of a coin.
      ii. The number of customers in a line in a minute.
      iii. If a bullet hits its target.

II. Binomial Distributions
   a. As the prefix "bi" suggests, binomial distributions deal with the number two. Each observation in such a distribution can be one of two results: success or failure.
      i. "Success" and "failure" is just nomenclature and does not suggest something good or bad happening. A "success" in a test for a disease can be a confirmation that the disease is present.
      ii. p is the probability of a success
      iii. q is the probability of a failure
      iv. Because there are only two outcomes, $p = 1 - q$.
      v. Examples: determining if a person needs or does not need corrective lens; testing if a peach is ripe or not; recording if a household has a pet or not.
   b. ***Binomial distributions assume the probability of success is constant.*** That is their defining assumption. This means each trial is independent (e.g. each customer has a 10% chance of redeeming a coupon).
      i. Another way to achieve independent trials is if you are replacing each selection after a trial (chances of pulling a poker chip from a bag and replacing it each time); or
   c. Mean

$$\mu = np$$

      i. n is the number of trials
      ii. p is the probability of success
   d. When we've determined probabilities, it's good to know how often you'll get three, four, or any other number of successes.

III. How Awesome Excel Is
   a. The most practical aspect of the binomial distribution distribution is the probability that a particular number of things will happen.
   b. But the equations for the various probability functions look, well, terrifying. The good news is that each of these equations are *already* in Excel.
   c. The main contribution you have is to know which equation to use. But first let's master how to make Excel tell you the result.
   d. There's no data file for this lesson; just open Excel.

IV. Binominal
   a. The command for binomial distribution is "=BINOM.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. *Number_s*: the number of successes (called x in the notes).
      ii. *Trials*: the number of attempts (called n in the notes).
      iii. *Probability_s*: the chance of success, expressed as a decimal (called p in the notes).
      iv. *Cumulative*: type in either a 1 or a 0 for this value. (Or type TRUE or FALSE.)
         1. A "0" (or FALSE) means Excel will tell you the probability of getting exactly x successes.
         2. A "1" (or TRUE) means Excel will tell you the probability of getting exactly x *or fewer* successes.
   b. Using Cumulative
      i. The cumulative function is very useful, especially since the chance of any number of successes is 1.
      ii. Want to know the chance of getting at least 2 successes but no more 6 successes? Find the probability of getting 6 or fewer and subtract off the probability of getting 1 or fewer.
      iii. Want to know the chance of getting 4 or more successes? Find the probability of getting 3 or fewer and subtract that value from 1.
   c. Suppose I distributed 1,000 coupons for my business and I know from past experience that each coupon has a 5% chance of being redeemed. How likely is it that exactly 60 coupons will be redeemed?
      i. First note that this is binomial: the chance of any coupon being redeemed is independent from other and there are only two options: the coupon will be redeemed or it won't be.
      ii. Type "=BINOM.DIST(60,1000,0.05,0)" and press ENTER.

iii. The result will be 0.01967, or just under a 2% chance.

d. The average number of redeemed coupons will be 50 (1,000 times 0.05); what is the chance 45 to 55 coupons will be redeemed?
   i. The BINOM.DIST.RANGE function helps with this. It lets you put in two different x values and will tell you the chance of getting between those values, inclusive. To find at least 45 coupons redeemed, but no more than 55 coupons redeemed, you'd put in: =BINOM.DIST.RANGE(1000,0.05,45,55)
   ii. The result is about 0.5752, or 57.52% chance.
   iii. Note we can get the same result by doing some subtraction: BINOM.DIST(55,1000,0.05,1)-BINOM.DIST(44,1000,0.05,1)
   iv. Why do we subtract off 44 instead of 45? Because we want to include 45 in the range. When the number of successes are "lumpy" like they are in discrete probability distributions, we have to be careful about such things.

V. Hypergeometric Distributions

a. In a binomial distribution, we discussed the assumption of each trial as independent.
b. What happens if the result of each trial affects the likelihood of success, we need to use a different discrete probability function: hypergeometric.
c. The mean is:

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

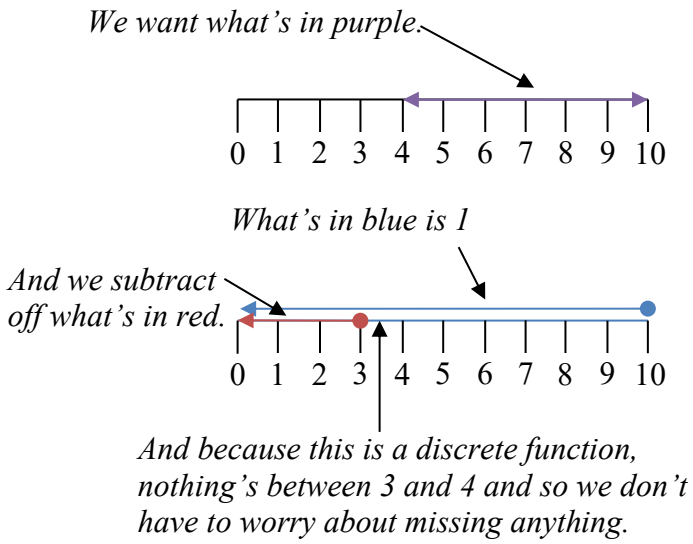
   i. Where *N* is the population size
   ii. *R* is the number of successes in the population
   iii. *n* is the sample size.

d. 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.)

e. Suppose ValuCorp's hiring engineers. They have 15 applicants, 5 of whom are women. Of these 15 applicants, 5 will be asked to be interviewed.
   i. We would expect 1 woman to be interviewed, but suppose no women are. Does this mean that there's sexism? We can use discrete probability distribution to find out.
   vi. First recognize that this is a hypergeometric distribution. For every person selected, the chance of selecting a woman changes, based on if the previous selection was a man or a woman.
   vii. Type =HYPGEOM.DIST(0,3,5,15,0) and press ENTER. You should get about 0.2637, or 26.37% chance. In other words, if you were selecting candidates at random, there's a pretty good chance you wouldn't select a woman. There's no evidence of sexism.
   viii. But what if 8 applicants were women? The probability of selecting only men falls to 0.7% chance. Of course, it could be that men happen to be more qualified than woman, but it's worth investigating.

f. Suppose you have a deck of cards and you draw 10 cards (without replacement). What is the chance you'll draw at least 4 hearts?
   i. Recall that there are 52 cards in a standard deck, 13 of which are hearts. How do you solve this problem?
   ii. You could find the probability of exactly 4 hearts, 5 hearts, etc. and then add them up but that takes too long. Instead, remember what we did before in binomial: we can subtract to find a range.
      1. Getting at least 4 hearts when you draw 10 is logically the same as getting at least 4 hearts, but no more than 10 hearts (because you can't get more than 10 hearts).
      2. So you would start with the chance of getting 10 or fewer hearts, but the chance of getting 10 or fewer hearts is 1 (because that covers all possibilities).
      3. So we would do this: =1- HYPGEOM.DIST(3,10,13,52,1)
   iii. Again, we have 3 instead of 4 in the Excel command because we want to exclude 3, 2, 1, and 0. We want to keep the scenario when we find 4 hearts.



VI. 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 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.)

e. 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. The question at the time was: could the Nazis aim these bombs?[1]
    i. Statistician R. D. Clarke realized that if the Nazis couldn't aim the bombs, the bombs would follow a Poisson distribution. He divided South London into 576 regions of equal areas, meaning an average of 0.9323 bombs per region. This is lambda.
    ii. 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 zero bombs?
    iii. Type "=POISSON.DIST(0,0.9323,0)" and press ENTER. You should get about 0.3936, or 39.36% chance.
    iv. Now do this for 1 bomb, 2 bombs, all the way up to 5 or more.
    v. Next, multiply each of these probabilities by 576; that gives you the expected number of regions that will get 0 bombs, 1 bomb, etc.
    vi. Here's the actual distribution; notice how close it is to the expected number of regions.

Number of bombs	*Expected number of regions*	*Actual number of regions*
0	226.7	229
1	211.4	211
2	98.5	93
3	30.6	35
4	7.1	7
5+	1.6	1

[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://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2019.01315.x