David Youngberg
BSAD 210—Montgomery College

LECTURE 25: DISCRETE PROBABILITY FUNCTIONS

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. Later we will discuss a *continuous probability distribution*: a distribution of data made up the results from a continuous random variable (which has outcomes of any numerical value or a large range of whole numbers).
   d. 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). Other ways to achieve independent trials:
      i. You are replacing each selection after a trial (chances of pulling a poker chip from a bag and replacing it each time); or
      ii. Your sample size is so small compared to the population, you don't affect the probability when you perform a trial. The

threshold for "big enough" is if your sample is less than 5% of your population (probability of finding money in any of 100 trash bags, selecting from the trash bags at the dump).

c. Mean

$$\mu = np$$

i. *n* is the number of trials
ii. *p* is the probability of success

d. Standard Deviation

$$\sigma = \sqrt{npq}$$

i. *n* is the number of trials
ii. *p* is the probability of success
iii. *q* is the probability of failure

e. 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 function 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. Try it for yourself. You should get 0.52896, or a 52.896% chance.

V. 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.

d. The standard deviation is:

$$\sigma = \sqrt{\frac{nR(N-R)}{N^2}}\sqrt{\frac{N-n}{N-1}}$$

VI. 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.

VII. Poisson Distribution

c. 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.

d. Requirements
   i. Mean must be the same for each interval.
   ii. Intervals cannot overlap.
   iii. Occurrences in each interval must be independent.

e. 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.

f. The standard deviation is:

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

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

VIII. Poisson

g. 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.)

h. During WWII, Germany launched a series of bombing raids on England, especially focusing on London. Between June 1944 and

March 1945, 535 flying-bombs fell on South London. At this time, South London was divided into 576 regions of equal areas, meaning an average of 0.929 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(2,0.929,0)" and press ENTER. You should get about 0.1704, or 17.04% chance.
iv. What is the chance that any particular region will suffer 3 or more bombs? You should get 0.0677, or about 6.77%.

[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.