MSPP PPOL 501–02 & -06: Fall 2014
Course Notes # 6: Discrete Random Variables; The Binomial Distribution
Professor Carolyn Hill

** Note: The notes from today draw on the following textbooks:

Runyon, Richard P., Kay A. Coleman, and David J. Pittenger. 2000. *Fundamentals of Behavioral Statistcs*, 9th Ed. (Boston: McGrawHill).

Weiss, Neil A. 2012. *Introductory Statistics*, 9th Ed. (Boston: Addison Wesley).

Wooldridge, Jeffrey M. 2003. *Introductory Econometrics: A Modern Approach*, 2nd Ed. Cincinnati: Thomson South-Western).

I. WHERE ARE WE?

- Moving from descriptive to inferential statistics
- Reviewing rules and "language" of probability as a precursor to inferential statistics

II. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS (GENERAL)

- A ***random variable*** is:

  "a quantitative variable whose value depends on chance" (Weiss, p. 212) (here, the "chance" refers to the method of selection)

  "[a variable] that takes on numerical values and has an outcome that is determined by a [procedure that can, at least in theory, be infinitely repeated, and has a well-defined set of outcomes]" (Wooldridge, p. 696)

- Examples: flips of a coin; dealing cards, free-throw shots, scores on tests, flight ticket holders who actually show up for flights, alcohol consumption of youth, ownership of cell phones, TV hours watched per week….the list goes on and on.

- Two kinds of random variables:

  1. **Discrete random variables** can take on a finite (or countably infinite) set of values. "possible values can be listed" (Weiss, p. 212)
     e.g.: number of children in family, flips of a coin, free throws

  2. **Continuous random variables** "can take on so many possible values that we cannot count them or match them up with the positive integers." (Wooldridge p 699)
     e.g.: price of a good; marathon finishing times

- Describing random variables:

  -- A random variable is often denoted by an italicized capital letter: E.g.:
  --- Let the number of children in a family be denoted by the random variable *X*.
  --- Let the price of milk be denoted by the random variable *Y*.

  -- A particular outcome, or manifestation, of a random variable is often denoted by a lower-case italicized letter. E.g.:
  --- In one family, $x = 3$ (e.g., two boys and one girl = 3 kids)

  --- The price of milk at the local grocery store today is y =\$3.09

  -- Note: in the above examples, think of these values being drawn at random from all possible values of each random variable, *X* and *Y*. I have just listed one possible outcome for each random variable: there are many other possible values

- We've already been working with random variables – we just haven't called them that. We've also already been working with the frequency distributions and **probability density functions** of random variables.

  -- The **probability distribution,** or **probability density function** (**pdf**) "summarizes the information concerning the possible outcomes of *X* and the corresponding probabilities" (Wooldridge, p. 698).

  -- The **cumulative distribution function (cdf)** is the sum of the probabilities for each event up to a certain event:

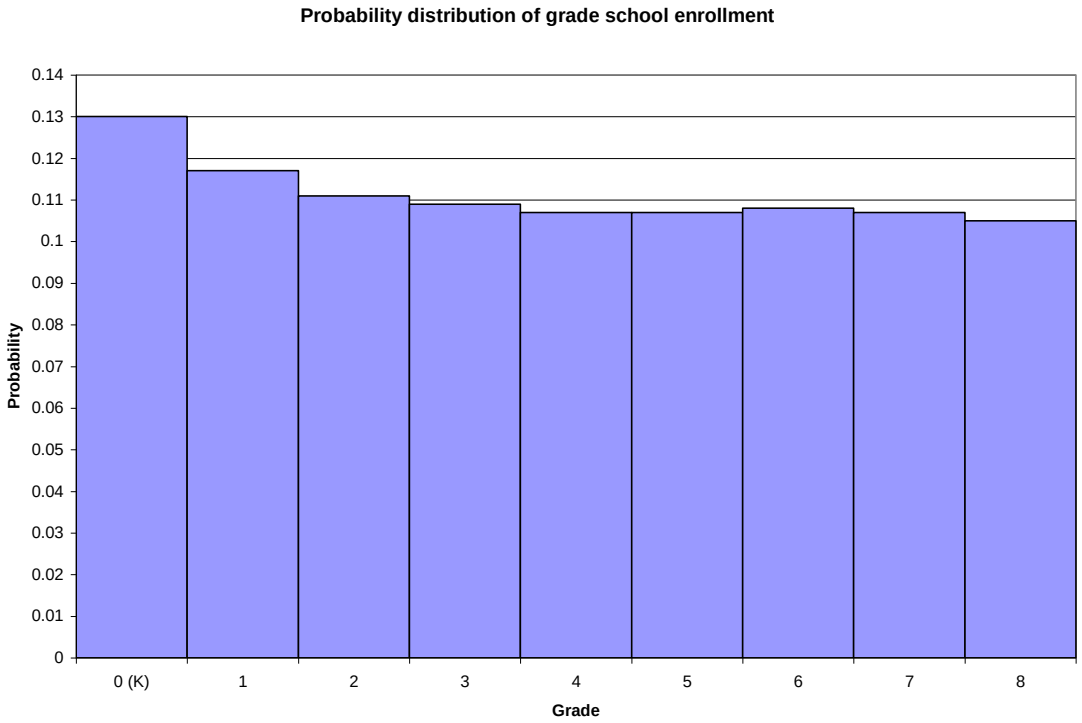
  E.g., distribution of enrollment in U.S. elementary schools:

Grade level x	*Frequency* f(x)	*Probability p(x)*	*Cumulative probability (cdf)*
0 (K)	4,208	0.130	0.130
1	3,769	0.117	0.247
2	3,596	0.111	0.358
3	3,518	0.109	0.467
4	3,447	0.107	0.574
5	3,447	0.107	0.681
6	3,486	0.108	0.789
7	3,457	0.107	0.896
8	3,398	0.105	1.001
TOTAL	32,326	1.001	n.a.

  --- The sum of probabilities from a pdf is always equal to 1 (i.e., all possible events and their probabilities have been accounted for) (the above sum is slightly greater than 1 due to the rounding of each of the individual probabilities)

--- Notation: the probability that random variable takes on a specific value is denoted as $P(X=k)$ = ___. For example: $P(X=2) = 0.111$.

-- A probability histogram shows the possible values of the random variables on the horizontal axis, and the probabilities associated with each value on the vertical axis. E.g.:



- So, you say to yourself, "Self, what's so different from these histograms and the ones that we looked in previous course notes when we were talking about descriptive statistics?"

-- The same basic idea is operating here: we're just pushing the concept a little bit.

--- In previous notes, we were using frequency tables and graphical depictions of data to **describe a set of data we had in hand**. It didn't matter whether those data were a random sample, a population of values, a convenience sample, few or many. We were simply using frequency distributions and graphics to describe what we had (and descriptive statistics will continue to serve this function).

--- Now we're combining these ideas with those of a population, sample, taking random samples from populations, drawing one or more observations at random from some set, random variables, etc.

--- In the current context, when we look at a table like the one above and think of the variable as a random variable, and selecting an observation or unit from the full set of data (however defined) at random, then the proportions in the data are probabilities – they constitute the pdf.

III. MEAN AND VARIANCE OF DISCRETE RANDOM VARIABLES

- Just as we talked about descriptive measures of central tendency and dispersion a few weeks ago, we also now need some way to describe the central tendency and dispersion for a random variable.

- For the mean of a random variable, the concept has a special name:

  "the **expected value (or expectation)** of *X*, denoted E(*X*) or sometimes μ_x or simply μ, is a weighted average of all possible values of *X*. The weights are determined by the probability density function. Sometimes, the expected value is called the *population mean*, especially when we want to emphasize that *X* represents some variable in a population" (Wooldridge, pp. 704-705).

- If *X* is a discrete random variable is a finite number of values, then the expected value is:

$$E(X) = \sum_{j=1}^{k} x_j \, p(x_j)$$

  E.g.: From elementary school example above:

Grade level x	*Frequency* f(x)	*Probability p(x)*	$x_j \, p(x_j)$
0 (K)	4,208	0.130	0
1	3,769	0.117	0.117
2	3,596	0.111	0.222
3	3,518	0.109	0.327
4	3,447	0.107	0.428
5	3,447	0.107	0.535
6	3,486	0.108	0.648
7	3,457	0.107	0.749
8	3,398	0.105	0.840
SUM			3.866

* For the dispersion of a discrete random variable, we measure the variance as:

$$\sigma^2 = \sum_{j=1}^{k} (X - \mu)^2 \, p(x_j) \text{ or as } \sigma^2 = \sum_{j=1}^{k} x_j^2 \, p(x_j) - \mu^2$$

  And the standard deviation is the square root of this value. For the example above, the standard deviation is 2.6.

IV. PROBABILITY DISTRIBUTION FOR DISCRETE SEQUENCES: THE SPECIAL CASE OF THE BINOMIAL DISTRIBUTION

- In the social sciences, we often deal with situations (or can transform a given situation into one of this type) where either something occurs or does not occur (or, something is the case or is not the case). E.g.:
  -- On a test: get the right answer or not
  -- A drug: is effective or not
  -- A person a certain number of years from his/her birth: is alive or not

- Defined this way, these kinds situations can be classified as discrete random variables, and in particular **Bernoulli random variables**, where

  $P(X=alive) = p$
  $P(X=notalive)=(1-p)$

  The outcome where the condition of interest is satisfied is called a "success" and the outcome where the condition of interest does not occur is called a "failure."

  Note that there's no normative meaning attached to "success" and "failure" – I could just as easily have defined the event of interest as "death." Then, if this "success" did not happen, the "failure" would be "not death." The important thing is to specify clearly how a "success" is defined in any given situation.

- Repeated, identical trials are Bernoulli trials if all these conditions hold:

  1. Each trial has 2 possible outcomes: success or failure (as above).
  2. The trials are independent.
  3. The probability of success is the same for each trial.

Example 1:

A multiple choice quiz has 5 questions. Four answers are possible for each question. Unprepared Student takes the quiz. What's the probability distribution for the number of answers he is expected to get right?

Possible Outcomes (x_j)	$p(x_j)$

- We could come up with these probabilities using the multiplication rule for independent events.

  * what's the probability of getting exactly 0 questions right on this quiz?

  e.g., what's the probability of getting exactly 2 questions right on this quiz?

  (a) First, figure out the probability of getting a particular two questions out of five questions right, e.g., Questions 1 and 2:

  (0.25*0.25*0.75*0.75*0.75) = 0.026367188

  (b) But, US might also have gotten 2 and 3 right (and missed 1, 4, and 5). The probability of that happening is also: (0.75*0.25*0.25*0.75*0.75) = 0.026367188.

  (c) Because the probability of getting 2 right and 3 wrong is going to be the same no matter which particular questions we're talking about, we need to figure out how many possible combinations of 2 questions right the student could have gotten.

  There are actually 10 possible combinations: think of a tree diagram. This list shows the combinations of possible question numbers that were possible to answer correctly:

  12; 13; 14; 15; 23; 24; 25; 34; 35; 45

  (d) So the probability of getting a specific combination correct (e.g. 1,2) is 0.026367188. Multiplied by 10 possible ways to get 2 correct, this gives us the probability of getting 2 correct of 0.26367188.

  (note: here's a web page (there are many) that will calculate the number of possible combinations for you: http://www.chemical-ecology.net/java/comb.htm) (note: this page will also calculate and show permutations, i.e., where order matters)

- We could fill out the probability distribution for this 0/1 random variable by going through the above process for each other possible type of outcome (i.e., getting 0 right, 1 right, 2 right, 3 right, 4 right, and 5 right).

- In the case like the one above, it's possible to compute these probabilities, but even gets a little mind-numbing in this simple example. What about more complicated

cases involving multiple possible trials, multiple possible outcomes, and thus multiple probabilities to compute?

- Help is on the way. If the 3 conditions above for Bernoulli trials hold, then the pdf has a special name and a special formula that simply implements the above logic in a briefer way. The **binomial distribution** is the probability distribution for the number of successes in a sequence of Bernoulli trials.

  Let X denote the total number of successes in n Bernoulli trials with success probability θ. The probability distribution of the random variable X is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \ for \ x = 0,1,2,...,n$$

- So, for the problem above…

$$P(X = 0) = \left(\frac{5!}{0!(5-0)!}\right)(0.25)^0 (0.75)^{5-0}$$

-- 0 right (0 successes):
$$= \left(\frac{5*4*3*2*1}{(1)*[5*4*3*2*1]}\right)(1)(0.237304688)$$
$$= \left(\frac{120}{120}\right)(0.237304688) = 0.237304688$$

$$P(X = 1) = \left(\frac{5!}{1!(5-1)!}\right)(0.25)^1 (0.75)^{5-1}$$

-- 1 right (1 success):
$$= \left(\frac{5*4*3*2*1}{(1)*[4*3*2*1]}\right)(0.25)(0.31640625)$$
$$= \left(\frac{120}{24}\right)(0.079101563) = 0.39551$$

$$P(X = 2) = \left(\frac{5!}{2!(5-2)!}\right)(0.25)^2 (0.75)^{5-2}$$

-- 2 right (2 successes):
$$= \left(\frac{5*4*3*2*1}{[2*1]*[3*2*1]}\right)(0.0625)(0.421875)$$
$$= \left(\frac{120}{12}\right)(0.026367188) = 0.263671875$$

-- etc.
$P(X=3) = 0.087891$
$P(X=4) = 0.014648$
$P(X=5) = 0.00097656$

Example 2:
The U.S. National Center for Health Statistics indicates that there is about an 80% chance that a person aged 20 will be alive at age 65. Suppose that 10 people aged 20 are selected at random. What's the probably that exactly 8 will be alive at age 65?

$$P(X = 8) = \left(\frac{10!}{8!(10-8)!} \right) (0.80)^8 (0.20)^{10-8}$$
$$= 0.30199$$

Did I calculate this out? No: you can do it using Excel, or some other program. You can also often find web tools that will calculate these distributions for you. An easy one that I found is: http://people.hofstra.edu/faculty/Stefan_Waner/RealWorld/stats/bernoulli.html

The key to using software like this is to know the values to input, and to be able to interpret the results.

- Note: other probability distributions for discrete variables exist (e.g., Poisson, discrete uniform, negative binomial, multinomial), but we won't cover them here.

V. MEAN AND VARIANCE OF A BINOMIAL RANDOM VARIABLE

- If X has a binomial distribution then the expected value is:

$$E(X) = \mu = np$$

E.g., from the multiple choice quiz example, the expected value of the distribution (i.e., the mean number of questions answered correctly) is:
$E(X) = \mu = 5 * 0.25 = 1.25$

The variance is:

$$Var(X) = \sigma^2 = np(1-p)$$

--- and the standard deviation is the square root of this value

E.g., the variance of questions answered correctly on the test above is:

$Var(X) = \sigma^2 = (5)(0.25)(.75) = 0.9375$, and the standard deviation is 0.968246.