

**POL 602**  
**Prof. Matthew Lebo**  
**Week 3 – September 13 and 15, 2005**  
**Discrete Random Variables**

Last time we discussed conditional probability, the additive and multiplicative laws, and Bayes' theorem which allows us to update our beliefs (Bayesian) or form posterior probabilities based on a prior and new information.

Here are 2 examples before we get to today's material.

Example 1: The search-and-rescue problem:

A plane has crashed somewhere in the mountains. As the leader of the SAR team, you have divided the area into three regions and believe that it is equiprobable for the plane to be in any of them. Your team searches region B (choosing randomly) and finds nothing; what is your updated probability that the plane is in region A?

$$\Pr(A | \bar{B}) = \frac{\Pr(\bar{B} | A) \Pr(A)}{\Pr(\bar{B} | A) \Pr(A) + \Pr(\bar{B} | \bar{A}) \Pr(\bar{A})}$$
$$\Pr(A | \bar{B}) = \frac{(1)(\frac{1}{3})}{(1)(\frac{1}{3}) + (\frac{1}{2})(\frac{2}{3})} = \frac{1}{2}$$

Where we want to know  $\Pr(A | \bar{B})$ , the probability the plane is in Area A given that it is not in area B.

So,  $\Pr(\bar{B} | A)$  is the probability the plane is not in area B given that its in area A &  $\Pr(\bar{B} | \bar{A})$  is the probability the plane is not in area B given that its not in area A.

Example 2: The "Monty Hall" problem:

Game show with three curtains. Behind one is a car, and gag gifts behind other two. Player picks one, then host shows her one of the other curtains at random is not the car. Should the player switch?

What is updated probability of staying? What about switching? How would Bayes' law work? Controversial problem, answer depends on the fact that Hall knows where the car is and is non-randomly showing you information (see Calvert paper on the preference for biased information).

Let's choose A

Then we are shown Curtain B.

What is the probability that the car is behind Curtain A?

Should we switch?

The *a priori* probability that the prize is behind door X =  $P(X) = 1/3$ .

The probability that Monty Hall opens door B if the prize is behind A,  $P(\text{Monty opens B} | A) = 1/2$ .

The probability that Monty Hall opens door B if the prize is behind B,  $P(\text{Monty opens B} | B) = 0$ .

The probability that Monty Hall opens door B if the prize is behind C,  $P(\text{Monty opens B} | C) = 1$ .

So, the probability that Monty Hall opens door B is then:

$$\begin{aligned} P(\text{Monty opens B}) &= P(A) * P(\text{M.O.B} | A) + P(B) * P(\text{M.O. B} | B) + P(C) * P(\text{M.O. B} | C) \\ &= 1/3 * 1/2 + 1/3 * 0 + 1/3 * 1 \\ &= 1/6 + 0 + 1/3 \\ &= 1/2 \end{aligned}$$

Then, by Bayes' Theorem,

$$P(A | \text{Monty opens B}) = \frac{P(\text{M.O.B} | A)P(A)}{P(\text{M.O.B})} = \frac{\frac{1}{2} * \frac{1}{3}}{\frac{1}{2}} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

and

$$P(C | \text{Monty opens B}) = \frac{P(\text{M.O.B} | C)P(C)}{P(\text{M.O.B})} = \frac{\frac{1}{3} * \frac{1}{3}}{\frac{1}{2}} = \frac{\frac{1}{9}}{\frac{1}{2}} = \frac{2}{9}$$

Here is a more extreme example which should make the concept easier for all to understand:

- \* Monty Hall presents you with 1,000,000 doors.
- \* You pick one.
- \* He goes through and opens 999,998 of the unselected doors (showing no prize), and leaves only one unselected door unopened.

With two doors remaining, does that leave you with a 50/50 chance of winning with either?

Clearly no. It is almost certain that Monty has deliberately avoided the one winning door in opening the other 999,998. You'd have a 99.9998% chance of winning if you choose to switch to the one he did not open.

To play, go to: [http://matti.usu.edu/nlvm/nav/frames\\_asid\\_117\\_g\\_1\\_t\\_5.html](http://matti.usu.edu/nlvm/nav/frames_asid_117_g_1_t_5.html)

### **Discrete Random Variables**

Recall that a random variable is a real-valued function defined over a sample space.

Today (and next time) we will talk only about *discrete* random variables.

A random variable  $Y$  is discrete if it can assume a finite or countably infinite number of distinct values. Countably infinite means integers, i.e. no decimals.

Examples: number of times a legislator wins elections, the result of a coin flip, number of amicus briefs filed in support of a Supreme Court case, etc.

Notation: Capital letters for random variable; lowercase letters for specific value it can take on (possible result of a given experiment), i.e. for the 6-sided die,

$$\forall P(Y = 1) = \frac{1}{6} \text{ or, in our book occasionally,}$$

$$p(1) = \frac{1}{6}$$

If we calculate all of the probabilities of a discrete random variable's possible outcomes, we have what is called a *probability distribution*.

This can be expressed graphically, in a table, or as a mathematical formula.

Example 3.1 (p.85): Foreman has 3 men and 3 women working and wants to pick 2 at random for a special job. Let  $Y$  denote the number of women chosen. Find the discrete probability distribution for  $Y$ .

$$\Pr(Y = 0) = \frac{\binom{3}{0}\binom{3}{2}}{\binom{6}{2}} = \frac{3}{15} = \frac{1}{5}$$

$$\Pr(Y = 1) = \frac{\binom{3}{1}\binom{3}{1}}{\binom{6}{2}} = \frac{9}{15} = \frac{3}{5}$$

$$\Pr(Y = 2) = \frac{\binom{3}{2}\binom{3}{0}}{\binom{6}{2}} = \frac{3}{15} = \frac{1}{5}$$

As noted above, we can thus display the distribution in three ways:

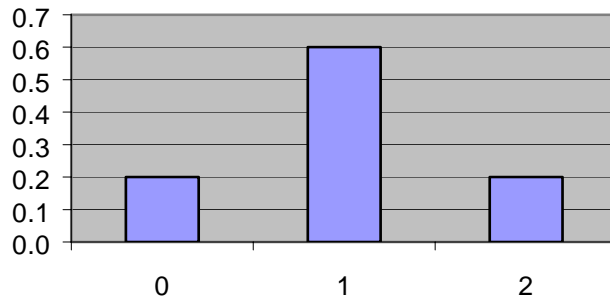
A) Table:

$y$	$p(y)$
<b>0</b>	1/5
<b>1</b>	3/5
<b>2</b>	1/5

B) Formula:

$$p(y) = \frac{\binom{3}{y}\binom{3}{2-y}}{\binom{6}{2}}, \quad y = 0, 1, 2$$

C) Graphically:



Note that, by the axiomatic definition of probability, two things must hold true for all discrete probability distributions:

$$1. \forall y, \quad 0 \leq p(y) \leq 1$$

$$2. \sum_y p(y) = 1$$

(note that  $\forall$  means “for all”)

### *Expectation*

Given this probability distribution, what is the mean number of women that we should expect the foreman to assign?

The mean of a probability distribution is known as the *expected value*.

This concept is one of the central points of this course.

Let  $Y$  be a discrete random variable with probability function  $p(y)$ . The expected value of  $Y$  is thus given by:

$$E(Y) = \sum_y yp(y)$$

For the previous example,

$$E(Y) = E(0) + E(1) + E(2)$$

$$E(Y) = 0\left(\frac{1}{5}\right) + 1\left(\frac{3}{5}\right) + 2\left(\frac{1}{5}\right)$$

$$E(Y) = 0 + \frac{3}{5} + \frac{2}{5} = 1$$

Another example: A stock trader is told that a given stock goes up on 90% of all days. When it does go up, it goes up by an average of 1 point. One days when the stock goes down, however, it goes down by 20 points on average. The stock trader wisely doesn't buy the stock or sells short, i.e. bets that the stock will go down.

Again,  $E(Y) = \sum_y yp(y)$ .

$$E(Y) = E(up) + E(down)$$

So,

$$= p(up) * increase + p(down) * decrease$$

$$= .9(1) + .1(-20)$$

$$= -1.1$$

And on any given day the trader can expect the stock to lose 1.1 points.

This example shows a key difference between *probability* and *expectation*.

Often, we are interested not only in the expected value of a random variable, but in the expected value of a function of a random variable.

If  $Y$  is a discrete random variable with probability function  $p(y)$  and  $g(y)$  is a real-valued function of  $Y$ , then the expected value of  $g(Y)$  is given by:

$$E[g(y)] = \sum_y g(y)p(y)$$

We can use this result to find the variance of a discrete prob. distribution:

$$V(Y) = E[(Y - \mu)^2]$$

$$V(Y) = \sum_y (y - \mu)^2 p(y)$$

Note: if the distribution is assumed to be an accurate characterization, we can further assume:

$$E(c) = c$$

$$E[cg(Y)] = cE[g(Y)]$$

$$E[g_1(Y) + g_2(Y) + \dots + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + \dots + E[g_k(Y)]$$

$$V(Y) = E(Y^2) - \mu^2$$

$$E(Y) = \mu, V(Y) = \sigma^2.$$

Example 3.2 (p.91):

Given the following distribution, find the mean, variance and std. dev. of  $Y$ :

$y$	$p(y)$
<b>0</b>	$\frac{1}{8}$

1	$1/4$
2	$3/8$
3	$1/4$

Remember that the variance of a random variable is the expected squared distance from the population mean.

$$\mu = E(Y) = \sum_y yp(y) = (0)(1/8) + (1)(1/4) + (2)(3/8) + (3)(1/4) = 1.75$$

$$\sigma^2 = E[(Y - \mu)^2] = \sum_y (y - \mu)^2 p(y)$$

$$= (0 - 1.75)^2(1/8) + (1 - 1.75)^2(1/4) + (2 - 1.75)^2(3/8) + (3 - 1.75)^2(1/4) = .9375$$

$$\sigma = +\sqrt{\sigma^2} = .97$$

Before moving on, we should state a few other useful properties of or results using expected values:

Where  $c$  is a constant:

$$E(c) = c$$

$$E[cg(Y)] = cE[g(Y)]$$

Also,

$$E[g_1(Y) + g_2(Y) + \dots + g_k(Y)] = E[g_1(Y)] + E[g_2(Y)] + \dots + E[g_k(Y)]$$

And,

$$V(Y) = E(Y^2) - \mu^2$$

The last result is a very useful calculational shortcut and is proven as follows:

$$\begin{aligned} V(Y) &= E[(Y - \mu)^2] \\ &= E(Y^2 - 2\mu Y + \mu^2) \\ &= E(Y^2) - E(2\mu Y) + E(\mu^2) \\ &= E(Y^2) - 2\mu E(Y) + \mu^2 \\ &= E(Y^2) - 2\mu^2 + \mu^2 \\ &= E(Y^2) - \mu^2 \end{aligned}$$

Note that this proof uses the other rules and thus serves as an excellent example. Note also that in the penultimate step, the definition of  $E(Y)=\mu$  is used.

Example 3.3 p. 94

Use this to find the variance of the random variable Y in the above example where the mean,  $\mu = 1.75$ .

$$E(Y^2) = \sum_y y^2 p(y) = (0)^2(1/8) + (1)^2(1/4) + (2)^2(3/8) + (3)^2(1/4) = 4$$

$$\text{Above formula gives us: } \sigma^2 = E(Y^2) - \mu^2 = 4 - (1.75)^2 = .9375$$

Two more results. First:

$$E(a + bx) = a + bE(x)$$

Now, the variance is a little more complicated, but you need to be able to manipulate these operators comfortably:

$$\begin{aligned} V(a + bx) &= E\left([a + bx - E(a + bx)]\right)^2 \\ &= E\left([a + bx - a + bE(x)]\right)^2 \\ &= E\left([bx - bE(x)]\right)^2 \\ &= E\left[b^2x^2 - b^2x(E(x)) + b^2(E(x))^2\right] \\ &= E\left[b^2x^2 - 2b^2\mu x + b^2\mu^2\right] \\ &= E\left[b^2(x^2 - 2\mu x + \mu^2)\right] \\ &= b^2E\left[(x - \mu)^2\right] \\ &= b^2V(x) \end{aligned}$$

Now we want to discuss two particular discrete probability distributions (first of several): the *Discrete Uniform or the Uniform on Integers*, and the *Binomial Distribution*.

### Uniform

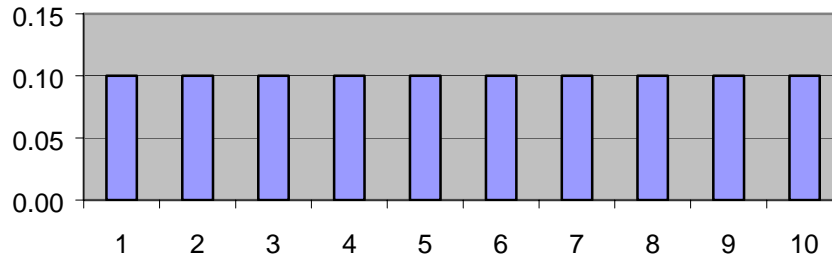
Suppose the value of a random variable Y is equally likely to be any of 1,2,3...k integers. A probability distribution is thus given by the function:

$$p(y) = \begin{cases} \frac{1}{k}, & y = 1, 2, \dots, k \\ 0 & \text{otherwise} \end{cases}$$

This is the uniform distribution on integers  $1, 2, \dots, k$  (or the discrete uniform distribution).

It represents the outcome of an experiment when we say that an integer  $[1, k]$  is chosen at random.

Graphically given  $k=10$ :



### Binomial

An experiment is said to be a binomial experiment if it consists of a sequence of identical and independent trials, each of which can result in only one of two outcomes (traditionally called *success* or *failure*).

Each of these trials can be referred to as a *Bernoulli* trial named for the Bernoulli family of statisticians.

We will let  $p$  be the probability of success, and  $q = (1-p)$  be the prob. of failure. The random variable of interest,  $Y$ , is the number of successes in  $n$  trials.

Example: An exam that is a series of 8 five-category multiple-choice questions. Not having studied, student  $x$  guesses randomly on each one. First, from last week's combinatorial problems, what is the probability of getting, say, any 6 of the eight questions right?

Every sample point in the compound event of interest will have 5 S's and 3 F's in it, such as  $S_1S_3S_5S_6S_7F_2F_4F_8$ . Since the trials (questions) are independent,  $p$  is the same for all questions ( $p = 1/5, q = 4/5$ ). Thus the probability of any single sample point is:

$$\frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{1}{5} \frac{4}{5} \frac{4}{5} \frac{4}{5}$$

or, more generally,

$$p^y q^{n-y}$$

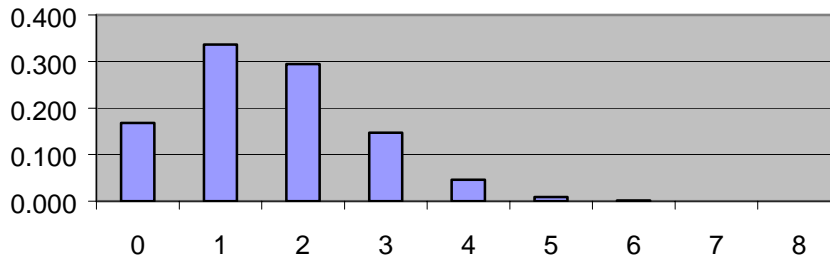
where  $y$  is the number of successes and  $n$  is the number of trials. But how many different ways can we get 6 questions right and 2 wrong? Since order doesn't matter, the answer is  $\binom{8}{6}$ , or, more generally,  $\binom{n}{y}$ . The probability of getting 6 of eight correct, then, is:

$$\binom{8}{6} \left(\frac{1}{5}\right)^6 \left(\frac{4}{5}\right)^2 \approx .00115$$

For every possible value of  $Y$ , what would the entire probability distribution look like?

	<b>p(y)</b>
0	<b>.168</b>
1	<b>.336</b>
2	<b>.294</b>
3	<b>.147</b>
4	<b>.046</b>
5	<b>.009</b>
6	<b>.0015</b>
7	<b>.00008</b>
8	<b>.00000256</b>

Graphically:



More generally, a random variable  $Y$  is said to have a binomial distribution based on  $n$  trials with success probability  $p$  if:

$$p(y) = \binom{n}{y} p^y q^{n-y}$$

Notice that this says that  $p(y) =$  The number of ways a sample point can be observed:  $\binom{n}{y}$  multiplied by the probability that each sample point is observed  $p^y q^{n-y}$ .

Note: <http://faculty.vassar.edu/lowry/binomialX.html> is an exact binomial calculator. This is a very cool site and worth trying out. There are also tables in the book, but they require interpolation and are not as useful.

The binomial distribution is useful in real world applications, including in political science where it is relevant to any dichotomous outcome over many trials (voting in a two-party election, answering yes or no survey questions, etc.).

Two additional useful results, which I will state, but not prove (I recommend looking at the book pp103-4 to examine these proofs): Given  $Y$  distributed binomial ( $Y \sim \text{binomial}$ )

$$\begin{aligned} \mu &= E(Y) = np \\ \sigma^2 &= V(Y) = npq \end{aligned}$$

Let's illustrate these with a new example:

Example: If the probability is .7 that any student with a 600 on the math portion of the GRE will get into graduate school, what is the probability that exactly 3 of 5 such students will get into graduate school?

$$p(3) = \binom{5}{3} (.7)^3 (.3)^2 = 0.31$$

What is the mean number (or expected value) of students with 600 math scores that are accepted?

$$E(Y) = np = (5)(.7) = 3.5$$

What is the variance of the binomial distribution in this example?

$$V(Y) = npq = (5)(.7)(.3) = 1.05$$

Recommended problems 3.4, 3.8, 3.10, 3.11, 3.17, 3.23(proof), 3.24, 3.25, 3.27, 3.28, 3.36, 3.37, 3.40, 3.41, 3.44, 3.45

## V. Discrete Random Variables II – Week 4 – September 20 and 22, 2005

Today we are going to continue to discuss the probability distributions of discrete random variables. Remember that what we are trying to get at here is a function that maps the observed values of a discrete random variable to the probability of observing those particular values.

Recall from last class that we introduced the binomial distribution (often called the Bernoulli distribution after its discoverer), a discrete probability distribution with the general form:

$$p(y) = \binom{n}{y} p^y q^{n-y}$$

where  $n$  is the total number of trials of a simple, two-outcome (dichotomous) experiment,  $y$  is the number of successful trials we are interested in (this is an observation on the random variable  $Y$ ),  $p$  is the probability of a “successful” trial, and  $q$  is  $(1-p)$ .

Remember also that for the binomial distribution:

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

$$\sigma^2 = V(Y) = npq$$

Note that the proofs of these equations are in the text, and are actually quite informative. If you had no problem following the discussion on expected values/expectation, then you should go ahead and work through them.

Let's illustrate how all this works again with another example:

Example: An investment analyst has tracked a certain blue-chip stock for the past six months and found that on any given day it either goes up a point or down a point. Furthermore, it went up on 25% of the days and down on 75%. What is the probability that at the close of trading four days from now the price of the stock will be the same as it is today? Assume that the daily fluctuations are independent events.

Let  $X$  = the number of days the stock rises.

Then  $X$  is binomial with  $n=4$  and  $p=0.25$ . The stock will be the same after four days only if  $X=2$ .

$$P(2) = \binom{4}{2} (.25)^2 (.75)^2 = 0.211$$

What is the mean number (or expected value) of days the stock goes up?

$$E(Y) = np = (4)(.25) = 1$$

What is the variance of the binomial distribution in this example?

$$V(Y) = npq = (4)(.25)(.75) = .75$$

Of course, many analysts who have followed similar strategies have quickly gone broke. It's a strategy only slightly smarter than my mother's "system" of betting on roulette numbers that have recently come up. It's not always appropriate to rely on history, or even recent history, to estimate the true value of  $p$  for your experiment.

### Geometric Probability Distribution

Next, I want to introduce three more discrete probability distributions. The first is the *Geometric* probability distribution.

Like the binomial distribution, the random variable with a geometric distribution also involves independent and identical trials, each of which can result in only two outcomes, success or failure.

The difference, in this case, is that we are interested not in the number of successes that occur, but the number of the trial on which the *first* success occurs.

So, the application would be a model of any process that can vary randomly in the number of trials before getting a first success, for example: the distribution of the number of times Pat Buchanan runs for President before actually getting elected?

How many days go by (assuming "discrete time") before my transmission breaks (i.e. the mean time before failure in an engineering context).

The sample space in an experiment like this is:

$$\begin{aligned} E1: & S \\ & \text{[Buchanan wins on the first election]} \\ E2: & F S \\ E3: & F F S \end{aligned}$$

$$E_k = \underbrace{FF \dots F}_{k-1} S$$

So,

$$P(Y = y) = P(E_y) = P(\underbrace{FF \dots F}_{y-1} S) = q^{y-1} p$$

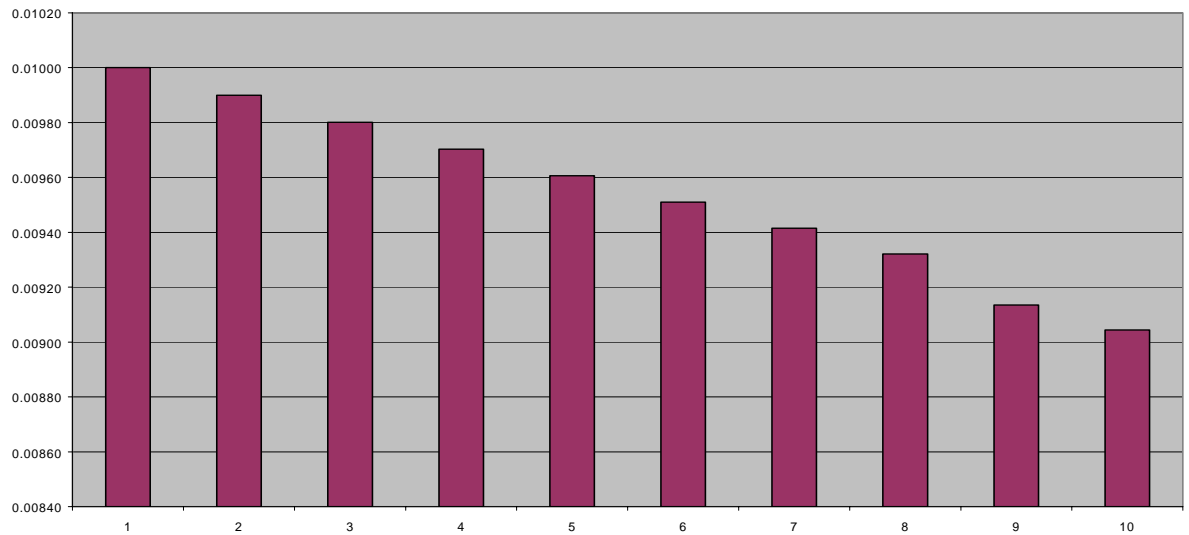
Thus, a random variable  $Y$  is said to have a geometric probability distribution if and only if:

$$p(y) = q^{y-1} p, \quad y = 1, 2, 3, \dots, \quad p \in [0, 1]$$

Example: what is the probability of Pat Buchanan's becoming President the tenth time he runs (assuming independent and identical elections which is of course impossible), given that we assume his chance of winning any given election is a constant .01?

$$P(Y = 10) = .99^{10-1} .01 \approx .00914$$

What does this distribution look like?



With values:

- 1 0.01000
- 2 0.00990
- 3 0.00980
- 4 0.00970
- 5 0.00961
- 6 0.00951
- 7 0.00942
- 8 0.00932
- 9 0.00914
- 10 0.00904

If the initial probability of winning is .01, why is the probability less than that for subsequent trials?

Note that the sum of these is .0954, not 1. The geometric distribution only works for a countably infinite random variable, and  $\sum_{y=1}^{\infty} p(y) = 1$ .

As in the case of the binomial distribution, I will provide you with the formulae for the mean and variance. Proofs are in the text:

$$\mu = E(Y) = \frac{1}{p}$$

$$\sigma^2 = V(Y) = \frac{1-p}{p^2}$$

The mean here is the expected number of trials it takes to have the first success.

As a slightly more realistic example than the Buchanan one, consider this: you are conducting a series of elite interviews of policy makers as part of a larger research project. It costs you a fixed amount (say \$10) in terms of phone charges and labor to attempt to interview each possible subject. In your grant proposal, how much money should you budget overall if you want to conduct 50 interviews and you think the probability of anyone actually responding to your request is 0.2?

First, what is the mean of the geometric distribution with  $p=.2$ ? This value is how many calls we should expect, on average, to have to make to reach a willing interviewee:

$$E(Y) = \frac{1}{p} = \frac{1}{.2} = 5$$

We should thus expect to have to make  $5(50) = 250$  calls to get our sample, thus we should budget at least \$2500 in our grant proposal.

What if we want to be more certain of getting our sample of 50 before we run out of money? By the empirical rule, if we are within 3 std. dev of the mean, we have about 99.7% certainty (actually it's not that simple) of getting the true value:

$$\sigma = \sqrt{V(Y)} = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{.8}{.04}} \approx 4.47$$

$$3\sigma = 13.41$$

So if we budget for  $14+5= 19$ ,  $19(50)= 950$  calls, for a total of \$9500, we are better than 99.7% likely to get our full sample of 50 without running out of funds.

This assumes that we are extremely unlucky in each and every one of our attempts to get a respondent.

### **Negative Binomial**

Lets quickly introduce the next two discrete probability distributions.

Both are frequently useful in our work. The next is the so-called *negative binomial distribution*.

Again, we are looking at independent and identical trials with one of two outcomes: success or failure.

The probability of success  $p$  stays the same from trial to trial.

The geometric distribution allows us to figure out when the first success occurs but what if we want to know the probability that a second or third or fourth success occurs on a given trial, e.g. that you get your second question wrong on the fifth question of a test?

If  $r$  is the number of success of interest (the  $r^{\text{th}}$ ),  $p$  is the probability of success and  $q$  is  $1-p$ , then:

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r}$$

Example:

A geological study indicates that an exploratory oil well drilled in a particular region should strike oil with  $p = .2$ . What is the probability that the third oil strike comes on the fifth well drilled? So,  $r = 3$  and  $y = 5$ :

$$P(Y = 5) = \binom{5-1}{3-1} .2^3 .8^2 = .0307$$

The formulae for the mean and variance (again without proofs) are:

$$\mu = E(Y) = \frac{r}{p}$$
$$\sigma^2 = V(Y) = \frac{r(1-p)}{p^2}$$

### Poisson

The next distribution is one that is of particular interest to our applied research. It is called the *Poisson* distribution (after its discoverer) and is used in modeling the probability that a particular event occurs a given number of times in a specified period.

For example, say we are interested in the number of bills reported out of the House Armed Services Committee in a single session of Congress. To model this experiment as a Poisson process, there are three assumptions we must make:

1. The number of occurrences in any two subintervals of time (dividing the session up into an arbitrary number of subintervals) must be independent. [This one is problematic in our example, and is actually problematic in many cases where this type of model is actually used]

2. The second condition is that the probability of an occurrence (a bill reported) in any short interval of time must be approximately proportional to the length of the time interval. In other words, as we cut the Congress up into shorter and shorter time intervals, the probability of a bill being reported in that time must decrease. This is reasonable in this case. (It would not be true if the probability varied over time, for example)

3. Lastly, we must be able to choose subintervals in which two occurrences cannot occur at the same time (this is a vast simplification of the real condition).

If these assumptions are met, we can say that our random variable of interest is distributed Poisson.

We derive this distribution from the binomial.

We have  $n$  subintervals of time and  $p$  is the probability of success in any one of those subintervals.

So this is a binomial experiment, but we don't know  $n$  and we don't know  $p$ .

If we let  $\lambda = np$  (number of trials times prob. of success) and take the limit of the binomial distribution function as  $n \rightarrow \infty$ , we get the Poisson distribution function:

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad \lambda > 0$$

Note that  $\lambda$  is lambda.

Recalling that the expected value of the binomial distribution is  $np$ , we should not be surprised that the mean of the Poisson is:

$$\mu = E(Y) = \lambda$$

What is a bit more interesting is that:

$$\sigma^2 = V(Y) = \lambda$$

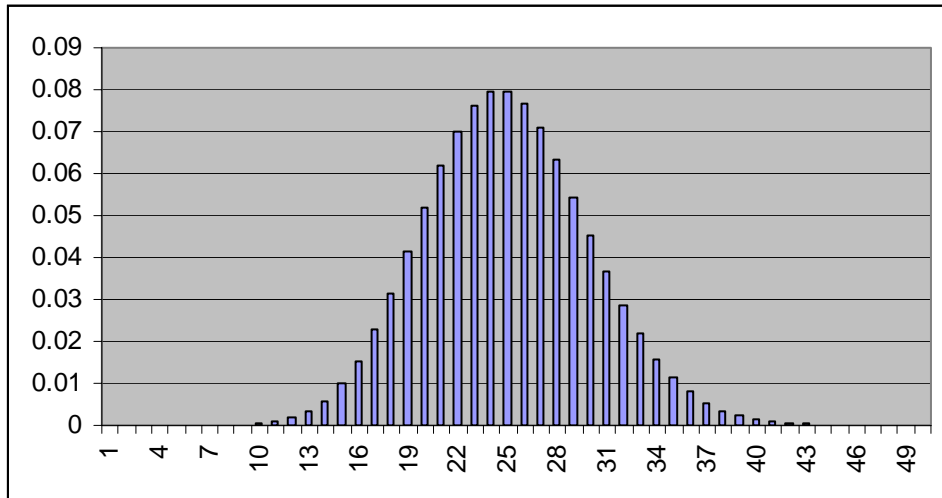
Definitely review the proofs for these results once you are comfortable with them (The variance proof is actually left as an exercise).

To finish our example, suppose that the historical average number of bills that the committee reports out in a session is 25 [Just a guess].

What is the probability that they only report out 10 bills?

$$P(Y = 10) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{25^{10}}{10!} e^{-25} \approx 0.00021$$

What does this probability distribution look like for, say, 0-50 bills out of committee?



Lastly, note that given the derivation of the Poisson distribution, it should not be surprising that when  $n$  is sufficiently large (say, greater than 100), the Poisson can be used as a computationally simpler approximation to the binomial or Bernoulli distribution: just let  $\lambda = np$ .

Tables 2 and 3 at the back of the book can be very handy for using the Poisson distribution. Spend some time going over them and trying to use them with examples.

Example II:

In its history, 1 in a hundred U.S. Senators have been Hispanic. Find the probability that a group of 50 Senators contains no Hispanics.

$$\lambda = (1) \frac{50}{100} = 0.5$$

$$P(Y=0) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{0.5^0}{0!} e^{-0.5} = \frac{e^{-0.5}}{0!} = 0.607 \quad P(y=5) = \frac{\binom{20}{5} \binom{60}{5}}{\binom{80}{10}} = 0.05$$

## Hypergeometric Distribution

Another useful distribution that will look familiar in a few minutes—this is actually a formalization of the method that we used to solve some combinatory problems.

Consider the case in which the sample size for a given discrete analysis is small relative to the population size.

For example (example 3.6): suppose that 40% of a large population of voters favor Jones. We take a random sample of 10 voters and we observe the number of voters who say they favor Jones.

If we let  $Y$  be a random variable denoting this number, how is it distributed?

A logical choice for modeling  $Y$  is the binomial distribution if we make a few assumptions:

1. The 10 people are approximately 10 identical, independent trials
2. There are only two outcomes—success (supporting Jones) or failure
3. The proportion of voters selecting Jones stays the same as I sample without replacement: .40 [THIS IS THE KEY ASSUMPTION].

What if 3 does not hold—that is, what if the conditional probability of success (voting for Jones) changes as individuals are drawn from the population? Equivalently, what if the sample size is large relative to the population size?

In other words, suppose that a finite population is composed of two types of elements, i.e. an urn filled with  $r$  red balls and  $b = N - r$  black balls. If we select a sample of size  $n$  and we are interested in  $Y$ , the number of balls in  $n$  that are red, then  $Y$  is said to have a *hypergeometric distribution*.

The formula for this distribution is straightforward and should look familiar—it's a generalization of the method we used to count sample points and solve questions such as the probability of the foreman selecting a team of exactly 2 women out of 3 people for the “special job.”

$$p(y) = \frac{\binom{r}{y} \binom{N-r}{n-y}}{\binom{N}{n}}$$

Example 3.16: Note that this is a classic problem (best of a set) for HR/personnel management. From a group of 20 Ph.D. Engineers, we select 10. What is the probability that the 10 selected include the best 5 of the group?

So,  $N=20$ ,  $n=10$ ,  $r=5$ . Remember  $r$  is the number in the population and  $y$  is the number we are interested in observing. Here, both equal 5.

We are interested in the probability of selecting all 5 of the 5 best:

$$p(5) = \frac{\binom{5}{5} \binom{15}{5}}{\binom{20}{10}} = \frac{\left(\frac{15!}{5!10!}\right)}{\left(\frac{20!}{10!10!}\right)} = \frac{21}{1292} = .0162$$

This is interesting: if we pick 10 out of 20 people at random, our chance of getting the 5 best of them is less than 2%, fairly counterintuitive.

And what is the probability of choosing 5 of the top 10 engineers when choosing 10 people at random?

I'm glad you asked, its: 
$$p(5) = \frac{\binom{10}{5} \binom{10}{5}}{\binom{20}{10}} = \frac{\left(\frac{10!}{5!5!}\right) \left(\frac{10!}{5!5!}\right)}{\left(\frac{20!}{10!10!}\right)} = \frac{63504}{184730.6} = .34376546$$

Here's a good website to do your factorials: <http://www.cs.uml.edu/~ytran/factorial.html>

Once again I will not prove the following results, but it holds that, for the hypergeometric distribution:

$$\mu = E(Y) = \frac{nr}{N}$$

$$\sigma^2 = V(Y) = n \left(\frac{r}{N}\right) \left(\frac{N-r}{N}\right) \left(\frac{N-n}{N-1}\right)$$

While this latter term looks complicated, we can also think of it in more familiar (binomial) terms:

Let  $p = \frac{r}{N}$  and  $q = 1 - p = \frac{N-r}{N}$ . Then the variance formula is:

$$\sigma^2 = npq \left(\frac{N-n}{N-1}\right)$$

which is the binomial distribution's variance multiplied by the last factor, which essentially corrects for the large sample size relative to the population. In fact, holding

the sample size constant,  $\lim_{N \rightarrow \infty} \left(\frac{N-n}{N-1}\right) = 1$  and this quantity can be ignored.

This quantity comes up again in survey research, where it is referred to as the *finite population correction factor*. When you work with survey data and you have sampled a large proportion of the population, you need this correction to properly account for variance. Failure to do so will cause an error in inference. We will return to this much later.

#### Example II

Keno is a very popular game to play in Las Vegas even though it ranks as one of the least favorable games in the sense that the odds are overwhelmingly in favor of the house. Still, it is often used as a way to continue gambling even while at the casino restaurants. A Keno card has 80 numbers, 1 through 80, from which the player selects a sample of size  $k$ , where  $k$  can be anything from 1 to 15. The “caller” then announces 20 winning numbers, chosen at random from the 80. If – and how much – a player wins depends on how many of his numbers match the 20 identified by the caller. Suppose that a player bets on a 10-spot ticket. What is his probability of getting 5 numbers right?

The size of the entire population is  $N = 80$ .

The number of numbers chosen is  $r = 20$ , leaving  $N - r = 60$  numbers not chosen.

We have a sample of  $n = 10$ , from which we want to know  $P(Y=y=5)$ .

$$P(y = 5) = \frac{\binom{20}{5} \binom{60}{5}}{\binom{80}{10}} = 0.05$$

Homework problems: 3.27, 3.29, 3.37, 3.45, 3.51, 3.55, 3.57, 3.59, 3.61, 3.74, 3.75, 3.77, 3.78, 3.85, 3.91, 3.93, 3.97, 3.101, 3.107