Discrete Random Variables
(Devore Chapter Three)

1016-351-01: Probability*

Fall 2011

Contents

Tuesday 20 September 2011

1 Random Variables

A random variable is a variable whose value is not definitively known, but is described probabilistically. Different values (or ranges of values) of the random variable are thought of as different events, with probabilities assigned to them. We use capital letters such as X or Y to refer to random variables, and lowercase letters to refer to specific values they can take. For example, we can consider the quantity $2+X$ and note that if $X = x$ then $2+X = 2+x$.

We will usually consider two kinds of random variables:

- *Discrete Random Variables* take on specific, separated values, and each possible value has a non-zero probability of occurring. The list of possible values can be finite (e.g., $\{1, 1.5, 2, 1.5, 3\}$) or "countably infinite", e.g., the set of all integers. Examples of discrete rvs are: the average of the numbers rolled on two dice, the number of children in a randomly selected family, etc.

- *Continuous Random Variables* can take on values in one or more intervals, with any value in the interval being possible. However, the probability of a continuous rv taking on any one specific value is zero. What is non-zero is the probability for the value of a continuous rv to lie within some range. Examples of continuous rvs are: the distance of a javelin throw, the height of a randomly selected child, the duration of a plane flight, etc.

For the time being (in Chapter 3) we will limit our attention to discrete rvs. We will return to continuous rvs in Chapter 4.

A very simple sort of discrete rv is a *Bernoulli random variable*. A Bernoulli rv can only take on the values 0 and 1.

1.1 Probability Mass Function

The behavior of a discrete random variable can be completely specified if we can assign a probability to each possible value. We can think of $X = x$, for a given x, as an event, and that event has a probability. We define something called the *Probability Mass Function*, or pmf:

$$p(x) = P(X = x) \tag{1.1}$$

For example for a Bernoulli random variable, $p(0) = P(X = 0)$ and $p(1) = P(X = 1)$. Note that in the case of a Bernoulli rv, the events $X = 0$ and $X = 1$ make up an exhaustive set of mutually exclusive alternatives, so

$$1 = P(X = 0) + P(X = 1) = p(0) + p(1) \qquad \text{for a Bernoulli rv} \tag{1.2}$$

More generally, since the total probability that X takes on *any* of its possible values is one,

$$\sum_x p(x) = 1 \tag{1.3}$$

This is known as a the *normalization condition* for the pmf of a discrete rv.

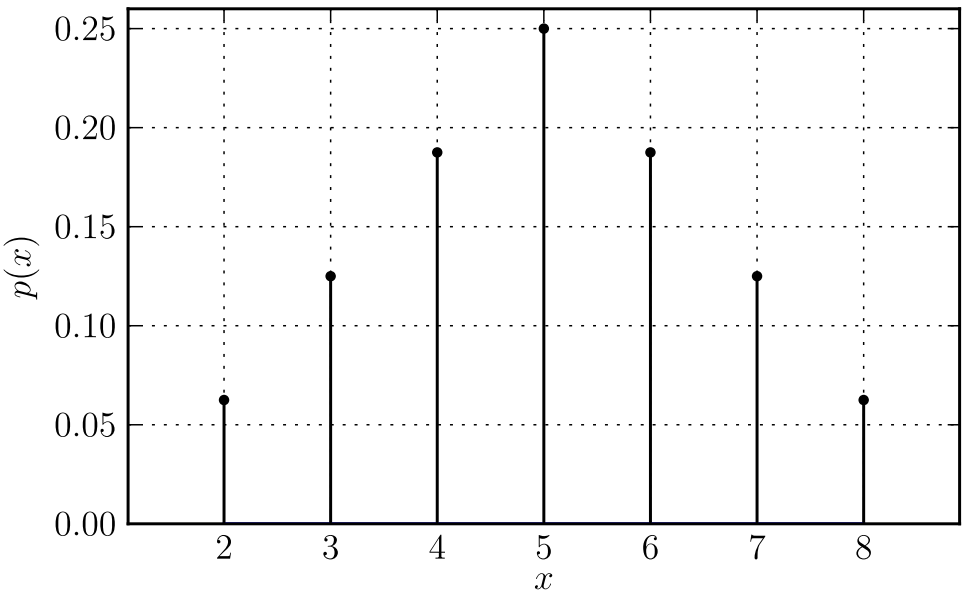
As a concrete example, suppose we roll two four-sided dice and define the rv X to be the sum of the numbers rolled. There are 16 possible outcomes, each of which is equally likely, and we can work out the pmf by just counting outcomes:

x	outcomes	$p(x)$
2	$(1,1)$	$1/16 = .0625$
3	$(1,2)$; $(2,1)$	$2/16 = .125$
4	$(1,3)$; $(2,2)$; $(3,1)$	$3/16 = .1875$
5	$(1,4)$; $(2,3)$; $(3,2)$; $(4,1)$	$4/16 = .25$
6	$(2,4)$; $(3,3)$; $(4,2)$	$3/16 = .1875$
7	$(3,4)$; $(4,3)$	$2/16 = .125$
8	$(4,4)$	$1/16 = .0625$

Note that we can check that this pmf is normalized:

$$\begin{aligned}\sum_x p(x) &= p(2) + p(3) + p(4) + p(5) + p(6) + p(7) + p(8) \\ &= .0625 + .125 + .1875 + .25 + .1875 + .125 + .0625 \\ &= 1\end{aligned} \tag{1.4}$$

We can also plot the pmf, drawing a stem to stress that it is non-zero only at the specified allowed values:



1.2 Cumulative Distribution Function

We often want to make statements about the probability for a rv to lie within a certain range. (In fact when we return to continuous rvs, those will be the only sorts of events that

we can assign probabilities to.) We can do this by adding the probabilities for all of the possible values that lie within that range. For instance, if we want the probability that X, the sum of two four-sided die rolls, is not more than 6 and not less than 4, that is

$$P(4 \leq X \leq 6) = \sum_{4 \leq x \leq 6} p(x) = p(4) + p(5) + p(6) = .1875 + .25 + .1875 = .625 \tag{1.5}$$

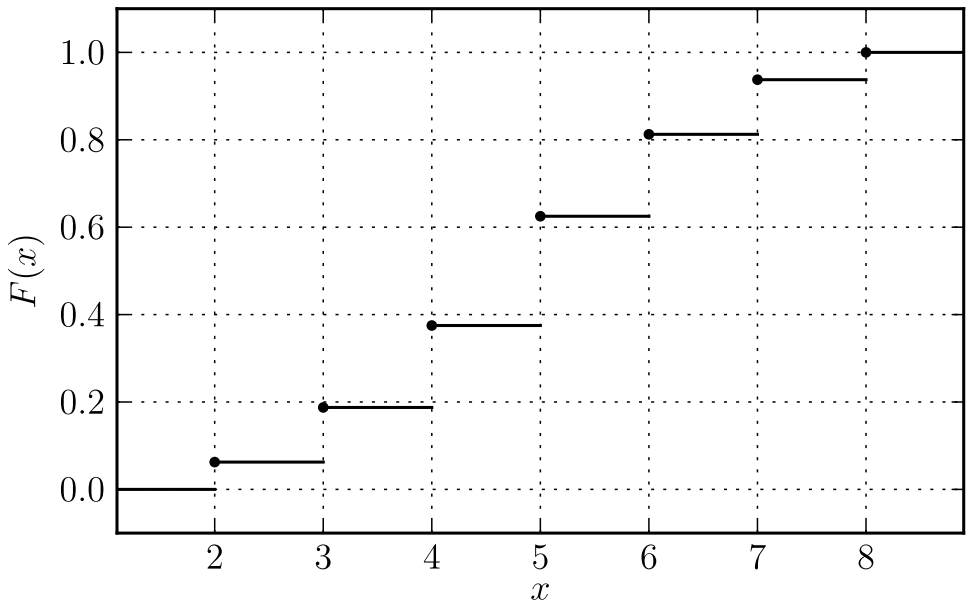
If we're going to do this a lot, though, the pmf is not so convenient to work with, since we have to keep adding up pmf values for every event we consider. It's more convenient to work with something called the *Cumulative Distribution Function*, which is just defined as a function of x, as the probability that X is less than or equal to x:

$$F(x) = P(X \leq x) = \sum_{y \leq x} p(y) \tag{1.6}$$

For example, in our 2d4 example, we get

$$F(x) = \begin{cases} 0 & x < 2 \\ .0625 & 2 \leq x < 3 \\ .1875 & 3 \leq x < 4 \\ .375 & 4 \leq x < 5 \\ .625 & 5 \leq x < 6 \\ .8125 & 6 \leq x < 7 \\ .9375 & 7 \leq x < 8 \\ 1 & 8 \leq x \end{cases} \tag{1.7}$$

Note that we'll always have $F(-\infty) = 0$ and $F(\infty) = 1$, if the original pmf is normalized. We can plot this cdf:



Note that the cdf is defined in between allowed values, and jumps discontinuously at those allowed values.

Once we've got the cdf, we can calculate the probability that X lies within any range by

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1) \tag{1.8}$$

We might need to be a little careful about our inequalities, though, since $x_1 \leq X$ is not the same as $x_1 < X$. E.g., returning to our 2d4 example,

$$P(4 \leq X \leq 6) = P(X \leq 6) - P(X < 4) \tag{1.9}$$

Now, we could find

$$P(X < 4) = P(X \leq 4) - P(X = 4) = F(4) - p(4) \tag{1.10}$$

but that sort of defeats the purpose of defining the cdf to get around the need to always add up values from the pmf. So a better strategy is to look at the cdf a little bit below 4, since there's no probability that e.g., $3.9 < X < 4$:

$$P(X < 4) = P(X \leq 3.9) + \cancelto{0}{P(3.9 < X < 4)} = F(3.9) \tag{1.11}$$

and so

$$P(4 \leq X \leq 6) = P(X \leq 6) - P(X < 4) = F(6) - F(3.9) = .8125 - .1875 = .625 \tag{1.12}$$

as before.

2 Expected Values

2.1 Mean of a Random Variable

A number of concepts from descriptive statistics translate into the realm of random variables. The first of this is the mean value. Consider for simplicity a Bernoulli random variable with $p(0) = .25$ and $p(1) = .75$. Although the simple average of 0 and 1 is .5, the random variable takes the value 1 75% the time and 0 only 25% of the time, so the appropriate mean value is .75.

In general, we can write the mean value as the weighted average of the possible values (weighted by the pmf)

$$\mu_X = \sum_x x\, p(x) \tag{2.1}$$

This is a special case of something called the *expected value*

$$E(X) = \sum_x x\, p(x) \tag{2.2}$$

Note that "expected value" doesn't literally mean that we expect the rv to take on exactly that value. The Bernoulli rv above can only take on the values 0 and 1, but the expected value is .75

Returning to our 2d4 example, the mean value of that rv is

$$\mu_X = E(X) = 2\,p(2) + 3\,p(3) + 4\,p(4) + \cdots + 8\,p(8) = 5 \tag{2.3}$$

Practice Problems

3.7, 3.9, 3.13, 3.21, 3.23, 3.25

Thursday 22 September 2011

2.2 Expected Value of a Function

For concreteness, let's consider a discrete rv with the pmf

x	.5	2	3
$p_X(x)$	.2	.5	.3

where we have written $p_X(x)$ rather than $p(x)$ to emphasize this is the pmf for the rv X. Recall that on Tuesday we defined the mean or expected value as

$$\mu_X = E(X) = \sum_x x\,p_X(x) \tag{2.4}$$

so in this case,

$$E(X) = (.2)(.5) + (.5)(2) + (.3)(3) = .1 + 1 + .9 = 2 \tag{2.5}$$

We can extend our definition of expected value to apply to functions of X, e.g., we can define $E(X^2)$ by considering $Y = X^2$ to be a new random variable, whose pmf we know:

y	.25	4	9
$p_Y(y) = P(X^2 = y)$	.2	.5	.3

so

$$E(X^2) = E(Y) \sum_y y\,p_Y(y) = (.2)(.25) + (.5)(4) + (.3)(9) = .05 + 2 + 2.7 = 4.75 \tag{2.6}$$

Of course, we'd also get the same result if we wrote down directly

$$E(X^2) = \sum_x x^2\,p_X(x) \tag{2.7}$$

without bothering to define a new rv. The general definition for the expected value of some function of a random variable is

$$E(h(X)) = \sum_x h(x)\,p(x) \tag{2.8}$$

(where we've now suppressed the subscript X and written $p_X(x)$ as $p(x)$ because there's only one rv to worry about.)

If the function is linear (or linear plus a constant), there's a result that simplifies the expected value a bit:

$$E(aX+b) = \sum_x (ax+b)\,p(x) = a\sum_x x\,p(x) + b\sum_x p(x) = aE(X)+b \tag{2.9}$$

Note that this only works for *linear* functions. For example, in general

$$E(X^2) \neq [E(X)]^2 \tag{2.10}$$

2.3 Variance of a Random Variable

Recall that the mean value μ_X of a random variable X is defined by analogy to a population mean, replacing the average over the population with a weighted average over the probability distribution, i.e., an expected value. Similarly, there is a quantity analogous to the population variance, which we define as

$$\sigma_X^2 = V(X) = E([X-\mu_X]^2) = \sum_x (x-\mu_X)^2 p(x) \tag{2.11}$$

In our simple example random variable, for which $\mu_X = 2$, we have

x	.5	2	3
$p(x)$	.2	.5	.3
$x-\mu_X$	-1.5	0	1
$(x-\mu_X)^2$	2.25	0	1

so

$$\begin{aligned}\sigma_X^2 = V(X) &= \sum_x (x-\mu_X)^2 p(x) = (2.25)(.2) + (0)(.5) + (1)(.3) \\ &= .45 + 0 + .3 = .75\end{aligned} \tag{2.12}$$

2.3.1 Standard Deviation

Of course, having written the variance $V(X)$ as σ_X^2, the natural extension is to call $\sigma_X = \sqrt{V(X)}$ the standard deviation of the rv. (In this case, $\sigma_X = \sqrt{.75} \approx .866$)

2.3.2 Shortcut Formula

If we apply the linearity property to the definition of the variance, we find

$$\begin{aligned}V(X) &= E([X-\mu_X]^2) = E(X^2 - 2\mu_X X + \mu_X^2) = E(X^2) + E(-2\mu_X X + \mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 = E(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= E(X^2) - [E(X)]^2\end{aligned} \tag{2.13}$$

This is just analogous to the shortcut formula we used to calculate the population (or sample) variance from the average (or sum) of the squares of the data values.

We can see this in action in our example, where we've already found $E(X) = 2$ and $E(X^2) = 4.75$, so the shortcut formula tells us

$$V(X) = E(X^2) - [E(X)]^2 = 4.75 - 2^2 = .75 \tag{2.14}$$

which is indeed just what we got by direct calculation.

3 Binomial and Related Distributions

3.1 The Binomial Distribution

We often think about an experiment with a binary outcome: success or failure. (In the language of random variables, the corresponding entity is a Bernoulli rv.) Suppose the probability of success is p. One of the definitions of probability is that if we could perform the experiment a large number of times under identical circumstances, the fraction of those trials which would give a successful result should tend towards p as the number of trials becomes large. But for a finite number of trials, the fraction will not be exactly p with complete certainty. If we flip a fair coin four times, there is a decent probability that we'll get something other than two heads and two tails. The *Binomial Distribution* gives the probability of getting some total number x of successes given a finite number n of trials, each of which has a probability p of success.

Example: suppose we are flipping an unbalanced coin with a probability of .6 to land on heads and .4 to land on tails, and we flip it twice. The different outcomes and their probabilities are:

Outcome	probability
HH	$(.6)(.6) = .36$
HT	$(.6)(.4) = .24$
TH	$(.4)(.6) = .24$
TT	$(.4)(.4) = .16$

Notice that the probability of getting a total of one head is not .24 but .48 because there are two different ways to do it:

# of heads	# of ways	prob of each way	total probability
2	1	.36	.36
1	2	.24	.48
0	1	.16	.16

In a general binomial experiment, the probability of getting a total of x successes and $n - x$ failures in n trials, with a probability of p for success and $1 - p$ for failure in each trial is

$$b(x; n, p) = (\#\text{ of ways})(\text{prob of each way}) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.1}$$

The probability for any particular sequence of x successes and $n-x$ failures is

$$p^x(1-p)^{n-x} \tag{3.2}$$

The number of distinct ways of getting x successes out of n trials is just "n choose x", which we learned last week:

$$\binom{n}{x} = \frac{n!}{(n-x)!x!} \tag{3.3}$$

So, putting it all together, with a bit of notation: if X is a binomial random variable, the total number of successes in n trials with a probability of success for each trial equal to p, we write

$$X \sim \text{Bin}(n,p) \tag{3.4}$$

and the pmf of X follows a binomial distribution

$$P(X=x) = p(x) = b(x;n,p) = \binom{n}{x} p^x(1-p)^{n-x} \quad x = 0,1,\ldots,n \tag{3.5}$$

The cumulative distribution function is

$$P(X \le x) = F(x) = B(x;n,p) = \sum_{y=0}^{x} b(y;n,p) \tag{3.6}$$

The mean and variance of a binomial distribution are

$$E(X) = np \tag{3.7a}$$

$$V(X) = np(1-p) \tag{3.7b}$$

It is possible to show both of these by manipulating the sums, but we prefer to argue that they make sense. Since p is the average long-term fraction of successes, it makes sense that $E(X/n) = E(X)/n$ should be equal to p. The variance is a little more complicated, but we can see that it's got some sensible features, e.g., it's zero if $p = 0$ (in which case every trial is a failure) or $p = 1$ (in which case every trial is a success). Also, note that the standard deviation $\sigma_X = \sqrt{np(1-p)}$ only grows like $\sqrt{n}$, so the expected standard deviation in the fraction of successes, σ_X/n, goes down as n gets large.

Practice Problems

3.29, 3.33, 3.43, 3.47 (use formula first, then verify w/table), 3.55, 3.59

Tuesday 27 September 2011

3.2 More on the Binomial Distribution

Recall: if we have a fixed number n of trials, each of which has a chance p of success, then if X is the random variable representing the total number of successes, X is called a binomial random variable, and we write

$$X \sim \text{Bin}(n,p) \tag{3.8}$$

and the pmf of X follows a binomial distribution

$$p(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n \tag{3.9}$$

because

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{3.10}$$

("n choose x") is the number of possible sequences of x successes and $n - x$ failures, and $p^x(1-p)^{n-x}$ is the probability of any one such sequence.

Also, the cdf of X is written

$$F(x) = B(x; n, p) = \sum_{y=0}^{x} b(y; n, p) \tag{3.11}$$

Now, $b(x; n, p)$ and $B(x; n, p)$ are explicitly defined, and you could in principle work them out by hand every time you need one. However, for all but the smallest values of x and $n-x$, that becomes tedious rather quickly. So this is the first of unfortunately many examples in probability and statistics where the standard approach is to look up the distribution in a statistical table. This is something of an anachronism, since these things can be easily calculated on a computer these days. But there is some middle ground between relying entirely on brute force and tables, and letting a statistical software package do your entire analysis for you. It's sort of like trigonometry: before the advent of calculators, you had to interpolate sines and cosines from values in a table. Now you can punch them into your calculator or call a function in a computer program. But you don't rely on a trigonometry package to work out all of your triangles for you. (And you hopefully remember the sines and cosines of a few simple angles like 0°, 30°, 45° and 90°.) So you shouldn't be ashamed to leave the calculating of something like $B(15; 35, .7)$ to a computer, but you should remember where it came from. (And also be able to write a little script in the language of your choice to calculate it.)

3.3 Hypergeometric Distribution

The binomial distribution is appropriate when the probability of each success or failure doesn't depend on the number of successes, like when calculating the probability of getting a total of three sixes when rolling ten six-sided dice. However, it doesn't give the right probabilities when you're taking elements out of a finite population. For example, consider the probability of getting exactly three spades in a five-card poker hand. Although 13 of the 52 cards in the deck are spades, we know from our example of calculating the probability of a flush that it's not just $b(3; 5, 1/4)$. Instead we need to calculate the number of different hands which have three spades and two non-spades, and divide it by the total number of possible poker hands.

- There are $\binom{13}{3} = \frac{13!}{3!10!}$ combinations of 3 out of the 13 spades.

- There are $\binom{39}{2} = \frac{39!}{2!37!}$ combinations of $5-3=2$ out of the $52-13=39$ non-spades.
- There are $\binom{52}{5} = \frac{52!}{5!47!}$ combinations of 5 out of the 52 cards.

So there are $\binom{13}{3}\binom{39}{2}$ different hands with exactly 3 spades, out of $\binom{52}{5}$ total hands, and the probability is

$$p(3\spadesuit) = \frac{\binom{13}{3}\binom{52-13}{5-3}}{\binom{52}{5}} \tag{3.12}$$

This is a special case of something called the *Hypergeometric Distribution*. If we randomly draw n items out of a set of N, M of which are successes, the number of successes we'll draw is a random variable X whose pmf is

$$h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{3.13}$$

We say that the n items are drawn "without replacement". If we drew cards one at a time and shuffled them back into the deck, we would be back in the situation where the probability of success for each card was the same no matter how many spades had already been drawn, and in that case we'd be back to a binomial distribution. It must also be the case that the hypergeometric distribution reduces to a binomial one when the number of draws is much less than the number of total objects, since in that case the probability of success on each draw shouldn't be affected much by the removal of a few successes or failures from the pool.

We can show this from the formula, that $h(x; n, M, N) \approx b(x; n, M/N)$ when $n \ll N$, $x \ll M$ and $n - x \ll N - M$. We first note that

$$\begin{aligned} \binom{N}{n} &= \frac{N!}{(N-n)!n!} = \frac{N(N-1)\cdots(N-n+1)}{n!} \\ &= \frac{N^n}{n!}\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right) \end{aligned} \tag{3.14}$$

In the limit of large N, each of the factors in parentheses goes to 1, so we can replace $\binom{N}{n}$ with $\frac{N^n}{n!}$. If we do the same thing with the other combinations, we get

$$\begin{aligned} h(x; n, M, N) &\approx \frac{\frac{M^x}{x!}\frac{(N-M)^{n-x}}{(n-x)!}}{\frac{N^n}{n!}} = \frac{n!}{x!(n-x)!}\frac{M^x(N-M)^{n-x}}{N^x N^{n-x}} \\ &= \binom{n}{x}\left(\frac{M}{N}\right)^x\left(\frac{N-M}{N}\right)^{n-x} = b\left(x; N, \frac{M}{N}\right) \end{aligned} . \tag{3.15}$$

3.4 Negative Binomial Distribution

Return now to a scenario where the per-trial chance of success is fixed. The binomial distribution is appropriate when the number of trials, n, is fixed ahead of time. If instead

we decide that we'll keep doing trials until we have a specified number of successes, r, the pmf for the number of failures, X, that we have in the meantime is $nb(x; r, p)$. Note that this is not n times $b(\cdots)$; the "nb" is taken as a unit. The total number of trials is $X + r$ (which is also a random variable).

This probability is a little more involved to estimate. To get $X = x$, we have to have $r - 1$ successes (and x failures) in the first $x + r - 1$ trials, and then a success in the last trial, so

$$\begin{aligned} nb(x; r, p) &= b(r-1; x+r-1, p) \cdot p = \binom{x+r-1}{r-1} p^{r-1}(1-p)^x p \\ &= \binom{x+r-1}{r-1} p^r (1-p)^x \end{aligned} \tag{3.16}$$

We can have any (non-negative integer) number of failures, so the pmf is defined for $x = 0, 1, 2, \ldots$.

3.5 Summary of Binomial and Related Distributions

The binomial, hypergeometric and negative binomial distributions are summarized in this table:

Distribution	Trials	Successes	Failures	Prob/Success
Binomial	n, fixed	X, rv	$n - X$, rv	p, fixed
Hypergeometric	n, fixed	X, rv	$n - X$, rv	drawn from M succ in pop of N
Negative Binomial	$X + r$, rv	r, fixed	X, rv	p, fixed

Distribution	pmf	Domain
Binomial	$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$	$0 \le x \le n$
Hypergeometric	$h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$\max(0, n - N + M) \le x \le \min(n, M)$
Negative Binomial	$nb(x; r, p) = \binom{x+r-1}{r-1} p^r (1-p)^x$	$0 \le x$

Practice Problems

3.65, 3.69, 3.71, 3.73, 3.75, 3.77

Thursday 29 September 2011

4 Poisson Distribution

Devore introduces the Poisson distribution starting with its pmf and then delving into the situations where it's relevant. Let's come at it from the other side. Consider the numerous statistical statements[1] you get like:

[1] I fabricated the last few significant figures in each of these numbers to produce a concrete example.

1. On average 118.3 people per day are killed in traffic accidents in the US
2. On average there are 367.2 gamma-ray bursts detected per year by orbiting satellites
3. During a rainstorm, an average of 929.4 raindrops falls on a square foot of ground each minute

In each of these cases, the number of events, X, occurring in one representative interval is a discrete random variable with a probability mass function. The average number of events occurring is a parameter of this distribution, which we sensibly write as μ.[2] From the information given above, we only know the mean value of the distribution, $E(X) = \mu$.

The key bit of extra information that makes the pmf a Poisson distribution is what happens if you break the interval up into smaller pieces. If each of the pieces can be treated as the same and independent of the others, the Poisson distribution is appropriate. So if we divide the day into 100 equal pieces of 14 minutes 24 seconds each, is the number of traffic deaths in each an independent random variable with a mean value of 11.83? Now, in practice this assumption will often not quite be true: the rate of traffic deaths is higher at some times of the day than others, some patches of ground may be more prone to be rained on because of wind patterns, etc, but we can imagine an idealized situation in which this subdivision works.

Okay, so how do we get the pmf? Take the interval in question (time interval, area of ground, or whatever) and divide it into n pieces. Each one of them will have an average of μ/n events. If we make n really big, so that $\mu/n \ll 1$, the probability of getting one event in that little piece will be small, and the probability that two or more of them happen to occur in the same piece is even smaller, and we can ignore it to a good approximation. (We can always make the approximation better by making n bigger.) That means that the number of events in that little piece, call it Y, has a pmf of

$$p(Y = 0) \approx 1 - p \tag{4.1a}$$
$$p(Y = 1) = p \tag{4.1b}$$
$$p(Y > 1) \approx 0 \tag{4.1c}$$

In order to have $E(Y) = \mu/n$, the probability of an event occurring in that little piece has to be $p = \mu/n$.

But we have now described the conditions for a binomial distribution! Each of the n tiny sub-pieces of the interval is like a trial, a piece with an event is a success, and a piece with no event is a failure. So the pmf for this Poisson random variable must be the limit of a binomial distribution as the number of trials gets large:

$$\begin{aligned} p(x;\mu) &= \lim_{n\to\infty} b(x;n,\mu/n) = \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x} \\ &= \frac{\mu^x}{x!} \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^n \frac{n!}{(n-x)!} \left(\frac{1/n}{1-\mu/n}\right)^x \end{aligned} \tag{4.2}$$

[2]Note that in the Seventh Edition of Devore this parameter was called λ, which led to all sorts of confusion later on. Calling it μ may be the most valuable change incorporated in the Eighth Edition.

Now, the ratio of factorials is a product of x things:

$$\frac{n!}{(n-x)!} = n(n-1)(n-2)\cdots(n-x+1) \tag{4.3}$$

The last factor is of course also the product of x things, i.e., x identical copies of

$$\frac{1/n}{1-\mu/n} = \frac{1}{n-\mu} \ . \tag{4.4}$$

But that means the two of them together give you

$$\frac{n!}{(n-x)!}\left(\frac{1/n}{1-\mu/n}\right)^x = \frac{n}{n-\mu}\frac{n-1}{n-\mu}\frac{n-2}{n-\mu}\cdots\frac{n-x+1}{n-\mu} \tag{4.5}$$

which is the product of x fractions, each of which goes to 1 as n goes to infinity, so we can lose that factor in the limit and get

$$p(x;\mu) = \frac{\mu^x}{x!}\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n = \frac{\mu^x}{x!}e^{-\mu} \ . \tag{4.6}$$

We've used the exponential function

$$e^\alpha = \lim_{n\to\infty}\left(1+\frac{\alpha}{n}\right)^n \ ; \tag{4.7}$$

This is the form one usually sees in the context of compound interest. [3] Here e is Euler's number, $e = 2.718\ldots$.

Note that in the pmf for a Poisson random variable X,

$$p(x;\mu) = \frac{\mu^x}{x!}e^{-\mu} \tag{4.8}$$

the most daunting part, the exponential, doesn't actually depend on x. It's a normalization constant which depends on the parameter μ. If we only care about the relative probabilities, we could write

$$p(x;\mu) = \mathcal{N}(\mu)\,\frac{\mu^x}{x!} \ ; \tag{4.9}$$

the fact that the constant $\mathcal{N}(\mu)$ is $e^{-\mu}$ is required by the normalization

$$\sum_{x=0}^{\infty} p(x;\mu) = \mathcal{N}(\mu)\sum_{x=0}^{\infty}\frac{\mu^x}{x!} = \mathcal{N}(\mu)\,e^\mu \ . \tag{4.10}$$

This uses the Taylor series

$$e^\mu = \sum_{x=0}^{\infty}\frac{\mu^x}{x!} \tag{4.11}$$

[3] If the rate times the term is α and we compound the interest at n equally spaced intervals during the term, the principal grows by a factor of $(1+\alpha/n)^n$; in the limit of continuously compounded interest, the principal doesn't grow infinitely, but only by a factor of e^α.

4.1 Example: Poisson distribution for $\mu = 2$

We can work out the pmf in this case. The normalization constant is $e^{-2} = .135335$. The pmf values are

$$p(0;2.) = \frac{2^0}{0!}e^{-2} = .135 \tag{4.12a}$$

$$p(1;2.) = \frac{2^1}{1!}e^{-2} = .271 \tag{4.12b}$$

$$p(2;2.) = \frac{2^2}{2!}e^{-2} = .271 \tag{4.12c}$$

$$p(3;2.) = \frac{2^3}{3!}e^{-2} = .180 \tag{4.12d}$$

$$p(4;2.) = \frac{2^4}{4!}e^{-2} = .0902 \tag{4.12e}$$

$$p(5;2.) = \frac{2^5}{5!}e^{-2} = .0361 \tag{4.12f}$$

$$p(6;2.) = \frac{2^6}{6!}e^{-2} = .0120 \tag{4.12g}$$

$$p(7;2.) = \frac{2^7}{7!}e^{-2} = .00344 \tag{4.12h}$$

$$p(8;2.) = \frac{2^8}{8!}e^{-2} = .000859 \tag{4.12i}$$

Note that $p(x;\mu)$ never goes to zero, but it gets very small. We can also say what the total probability of being above a certain point is, e.g.,

$$\begin{aligned} P(X > 4) &= 1 - P(X \le 4) = 1 - e^{-2}\left(1 + 2 + 2 + \frac{4}{3} + \frac{2}{3}\right) = 1 - 7e^{-2} \\ &= 1 - .947 = .053 \end{aligned} \tag{4.13}$$

Practice Problems

3.79, 3.81, 3.83, 3.85, 3.91, 3.93