Lecture 7

The Five Basic Discrete Random Variables

1. Binomial
2. Hypergeometric
3. Geometric
4. Negative Binomial
5. Poisson

Remark

On the handout "The basic probability distributions" there are six distributions. I did not list the Bernoulli distribution above because it is too simple.

In this lecture we will do 1. and 2. above.

The Binomial Distribution

Suppose we have a Bernoulli experiment with $P(S) = P$, for example, a weighted coin with $P(H) = P$. As usual we put $q = 1 - p$.
Repeat the experiment (flip the coin). Let $X = \sharp$ of successes ($\sharp$ of heads).
We want to compute the probability distribution of X. Note, we did the special case $n = 3$ in Lecture 6, pages 4 and 5.

Clearly the set of possible values for X is $0, 1, 2, 3, \ldots, n$.
Also

$$P(X = 0) = P(TT \quad T) = qq \ldots q = q$$

Explanation

Here we assume the outcomes of each of the repeated experiments are independent so

$$P((T \text{ on } 1^{\text{st}}) \cap (T \text{ on } 2^{\text{nd}}) \cap \cdots \cap (T \text{ on} n\text{-th})$$

$$P(T \text{ on } 1^{\text{st}}) P(T \text{ on } 2^{\text{rd}}) \ldots P(T \text{ on } n\text{-th})$$

$$q\, q \ldots q = q^n$$

Note T on 2^{nd} mean) T on 2^{nd} with no other information so

$$P(T \text{ on } 2^{\text{nd}}) = q.$$

Also

$$P(X = n) = P(HH \dots H) = P^n$$

Now we have to work
What is $P(X = 1)$?

Another standard mistake

The events $(X = 1)$ and $\underbrace{HTT \dots T}_{n-1}$ are NOT equal.

Why - the head doesn't have to come on the first toss

So in fact

$$(X = 1) = HTT \dots T \cup THT \dots T \cup \cdots \cup TTT \dots TH$$

All of the n events on the right have the same probability namely pq^{n-1} and they are mutually exclusive. There are n of them so

$$P(X = 1) = npq^{n-1}$$

Similarly

$$P(X = n - 1) = npq^{n-1}$$

(exchange H and T above)

The general formula

Now we want $P(X = k)$

First we note

$$P(\underbrace{H \dots H}_{k} \ \underbrace{TT \dots T}_{n-k}) = p^k q^{n-k}$$

But again the heads don't have to come first. So we need to

(1) Count all the words of length n in H and T that involve k. It's and $n - k$ T's.

(2) Multiply the number in (1) by $p^k q^{n-k}$.

So how do we solve 1. Think of filling n slot's with the H's and $n - k$ T's

$$\underbrace{\_\ \_\ \_\ \_\ \_\ \_\ \_}$$

Main Point

Once you decide where the kH's go you have no choice with the T's. They have to go in the remaining $n - k$ slots.
So choose the k-slots when the heads go. So we here to make a choose of k things from n things so $\binom{n}{k}$.

So,

$$P(X = k) = \binom{n}{k} P^k q^{n-k}$$

So we have motivated the following definition.

Definition

A discrete random variable X is said to have binomial distribution with parameters n and p (abbreviated $X \sim \mathrm{Bin}(n, p)$)
If X takes values $0, 1, 2, \ldots, n$ and

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, 0 \leq k \leq n. \qquad (*)$$

Remark

The text uses x instead of k for the independent (i.e., input) variable. So this would be written

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

I like to save x for the case of continuous random variables.

Finally we may write

$$P(k) = \binom{n}{k} p^k q^{n-k}, \quad 0 \leq k \leq n \qquad (**)$$

The text uses $b(\cdot, n, p)$ for $p(\cdot)$ so would write for (**)

$$b(k, n, p) = \binom{n}{k} p^k q^{n-k}$$

The Expected Value and Variance of a Binomial Random Variable

Proposition

Suppose $X \sim \mathrm{Bin}(n, p)$*. Then* $E(X) = np$ *and* $V(X) = npq$ *so* σ *= standard deviation* $= \sqrt{npq}$*.*

Remark

The formula for $E(X)$ *is what you might expect. If you toss a fair coin 100 times the* $E(X) =$ *expected number of heads* $np = (100)\left(\frac{1}{2}\right) = 50$*.*

However if you toss it 51 times then $E(X) = \frac{51}{2}$ *- not what you "expect".*

Using the binomial tables

Table A1 in the text
pg. 664-666 tabulates the cdf $B(x, n, p)$ for $n = 5, 10, 15, 20, 25$ and selected values of p.

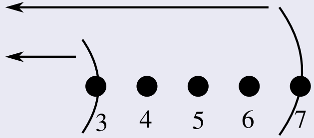
Example (3.32)

Suppose that 20% of all copies of a particular text book fail a certain binding strength text. Let X denote the number among 15 randomly selected copies that fail the test. Find

$$P(4 \leq X \leq 7).$$

Solution

$X \sim \mathrm{Bin}(15, .2)$. *We wont to compute* $P(4 \le X \le 7)$ *using the table on page 664. So how to we write* $P(4 = leqX \le 7)$ *in terms of terms of the form* $P(X \le a)$



Answer

$$(\sharp) P(4 \le X \le 7) = P(X \le 7) - P(X \le 3)$$

So

$$P(4 \le X \le 7) = B(7, .15, .2) - B(3, .15, .2)$$

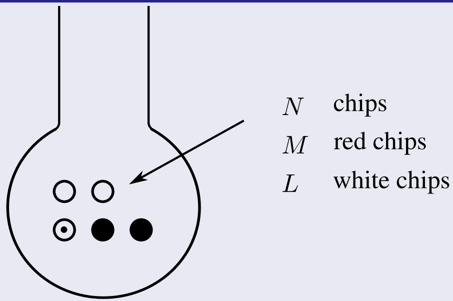
from table

$$= .996 - .648$$

N.B. *Understand* ($\sharp$)*. This the key using computers and statistical calculators to compute.*

The hypergeometric distribution

Example



Consider an urn containing N
Chips of which M are red and $L = N - M$ are white. Suppose we remove n chips without replacement so $n \leq N$.
Define a random variable X by $X = \sharp$ of red chips we get.

Find the probability distribution of X.

Proposition

$$P(X = k) = \frac{\binom{M}{k}\binom{L}{n-k}}{\binom{N}{n}} \qquad (*)$$

if

$$(b) \qquad \underbrace{\max(0, n-L) \le k \ne \min(n, M)}$$

This means $k \le$ both n and M and both 0 and $n - L \le k$.
These are the possible values of k, that is, if k doesn't satisfy (b) then

$$P(X = k) = 0.$$

Proof of the formula (*)

Suppose we first consider the special case where all the chips are red so

$$P(X = n).$$

This is the same problem as the one of finding all hearts in bridge.

$$\text{red chip} \longleftrightarrow \text{heart}$$
$$\text{white chip} \longleftrightarrow \text{non heart}$$

So we use the principle of restricted choise

$$P(X = n) = \frac{\binom{M}{n}}{\binom{N}{n}}$$

This agrees with (*).

But (*) is harder because we have to consider the case where there are kMn red chip. So we have to choose $n - k$ white chips as well.

So choose k red chips $-\binom{M}{k}$ ways, then for each such choice, choose $n - k$ white chipts $\binom{L}{n-k}$ ways.

So

$$\sharp \left\{ \begin{array}{l} \text{choices of } \underline{\text{exceptly}} \\ k \text{ red chips} \\ \text{in the } n \text{ chips} \end{array} \right\} = \binom{M}{k}\binom{L}{n-k}$$

Clearly there are $\binom{N}{n}$ ways of choosing n chips from N chips so (*) follows.

Definition

If X is a discrete random variable with pmf defined by page 14 then X is said to have hyper geometric distribution with parameters n, M, N. In the text the pmf is denoted

$$h(x; n, M, N).$$

What about the conditions

$$\max(0, n - L) \leq k \leq \min(n, M) \tag{b}$$

This really means

$$k \leq \text{ both } n \text{ and } M \tag{b_1}$$

and

$$\text{both } 0 \text{ and } n - L \leq k \tag{b_2}$$

(b_1) says

$k \leq n \quad \longleftrightarrow \quad$ we can't choose more then n red chips because we are only choosing n chips in to to

$k \leq M \quad \longleftrightarrow \quad$ because there are only M red chips to choose from

(b_2)

$$k \geq 0 \text{ is obvious.}$$

So the above three inequalities are necessary. At first glance they look sufficient because if k satisfies the above three inequalities you can certainly go ahead and choose k red chips.
But what about the white chips? We aren't tone yet, you have to choose $n - k$ white chips and there are only L white chips available so if $n - k > L$ we are sun k.
So we must have

$$n - k \leq L \Leftrightarrow k \geq n - L$$

This is the second inequality of (b_2). If it is satisfied we can go ahead and choose the $n - b$ white chips to the inequalities in (b) are necessary and sufficient.

Proposition

Suppose X has hypergeometric distribution with parameters n, M, N. Then

(i) $E(X) = n\dfrac{M}{N}$

(ii) $V(X) = \left(\dfrac{N-n}{N-1}\right) n\dfrac{M}{N}\left(1 - \dfrac{M}{N}\right)$

If you put

$$P_1 = \frac{M}{N} = \begin{array}{l}\textit{the probability of getting}\\ \textit{a red disk on the first draw}\end{array}$$

then we may rewrite the above formulas as

$$\left.\begin{aligned} E(X) &= nP \\ V(X) &= \left(\frac{N-n}{N-1}\right) npq \end{aligned}\right\} \begin{array}{l}\textit{reminiscent}\\ \textit{of the}\\ \textit{binomial}\\ \textit{distribution}\end{array}$$

Another way to Derive (*)

There is another way to derive (*) - the way we derived the binomial distribution. It is way harder.

Example

Take $n = 2$

$$P(X=0) = \frac{L}{N}\frac{L-1}{N+1}$$

$$P(X=2) = \frac{M}{N}\frac{M-1}{N-1}$$

$$\begin{aligned} P(X=1) &= P(RW) + P(WR) \\ &= \frac{M}{N}\frac{L}{N-1} + \frac{L}{N}\frac{M}{N-1} \\ &= 2\frac{M}{N}\frac{L}{N-1} \\ &= 2\frac{M}{N}\frac{L}{N-1} \end{aligned}$$

In general, we claim that all the words with kR's and $n - k$ W's have the some probability. Indeed each of these probabilities are fractions with the same denominator

$$N(N-1)\dots(N-n-1)$$

and they have the same factors in the numerator scrambled up

$$M(M-1)(M-L+1) \quad \text{and} \quad L(L-1),\dots,(L-n-k+i)$$

But the order of the factors doesn't matter so

$$\begin{aligned} P(X=k) &= \binom{n}{k} P(\overbrace{R\dots R}^{k} W\dots W) \\ &= \binom{n}{k} \frac{M(M-1)\dots(M-k+1)L(L-1)\dots(L-n-k+1)}{N(N-1)\dots N(-n+1)} \end{aligned}$$

Why is (*) equal to this?

$$(*) = \frac{\binom{M}{k}\binom{L}{n-k}}{\binom{N}{n}}$$

cancelling $(M-k)!$

cancelling $(L-n-k+1)$

$$= \frac{\frac{M(M-1)\ldots(M-k+1)}{k!}\ \frac{L(L-1)\ldots(L-n-k+1)}{(n-k)!}}{\frac{N(N-1)\ldots(N-n+1)}{n!}}$$

goes on top

$$(*) = \frac{\binom{M}{k}\binom{L}{n-k}}{\binom{N}{n}}$$

$$= \frac{\frac{M(M-1)\ldots(M-k+1)}{k!}\ \frac{L(L-1)\ldots(L-n-k+1)}{(n-k)!}}{\frac{N(N-1)\ldots(N-n+1)}{n!}}$$

exercise in fractions

$$= \frac{n!}{k!(n-k)!}\frac{M(M-1)\ldots(M-k+1)L(L-1)\ldots(L-n-k+1)}{N(N-1)\ldots(N-n+1)}$$

$$= \binom{n}{k}\frac{M(M-1)\ldots(M-k+1)L(L-1)\ldots(L-n-k+1)}{N(N-1)\ldots(N-n+1)}$$

Obviously, the first way (*) is easier so if you are doing a reel-world problem and you start getting things that look like (**) step back and see if you can use the first method instead. You will tend to try the second method first. I will test you on this later.

Prediction (I was wrong before)
Most of you will use the second (wrong) method.

An Important General Problem

Suppose you draw n chips <u>with replacement</u> and let X be the number of red chips you get. What distribution does X have?

This explains (a little) the formulas on page 21. Note that if N is far bigger than n then it is almost like drawing with replacement. “The urn doesn’t notice that any chaps have been removed because so few (relatively) have been removed.”

In this case

$$\frac{N-n}{N-1} = \frac{N\left(1-\frac{n}{N}\right)}{N\left(1-\frac{1}{N}\right)} \approx \frac{N}{N} = 1$$

(because N is huge $\frac{1}{N}$ and $\frac{n}{N} = 0$)

So $V(X) \approx npq$ (see the bottom at pg. 21)

This is what is going on in page 118 of the text.

The number $\frac{N-n}{N-1}$ is called the "finite population correction factor".