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.

1. 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 $X = \#$ of success ($\#$ of heads). We want to compute the probability distribution of X. Note, we did the special case $n = 3$ in Lecture 6, pgs. 4 & 5. Clearly, the set of possible values for X is $0, 1, 2, 3, \cdots, n$. Also,

$$P(X = 0) = P(TTT) = qq \cdots q = q^n.$$

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

$$\begin{aligned} &P((T \text{ on } 1^{st}) \cap (T \text{ on } 2^{nd}) \cap \cdots \cap (T \text{ on } \ n-th)) \\ &\quad = P(T \text{ on } 1^{st}) P(T \text{ on } 2^{nd}) \cdots P(T \text{ on } \ n-th) \\ &\quad = qq \cdots q = q^n. \end{aligned}$$

Note: T on 2^{nd} means T on 2^{nd} *with no other information* so

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

Also,

$$P(X = n) = P(HH \cdots H) = p^n.$$

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

Another standard mistake. The events $(X = 1)$ and $\underbrace{HTT\cdots T}$ are *NOT* equal. Why – *the head doesn't have to come on the first toss.* So in fact

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

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

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

Similarly,

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

(exchange H and T above).

The general formula. Now we want $P(X = k)$. First we note

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

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

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

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

Main Point. Once you decide where the $k\ H's$ 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 have to make a choice 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 value $0, 1, 2, \cdots, n$ and*

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

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

$$p(k) = \binom{n}{k} p^k q^{n-k}, \quad 0 \le k \le n. \tag{**}$$

The text uses $b(\cdot; n, p)$ for $p(\cdot)$ so we 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 $X \sim \text{Bin}(n, p)$. Then $E(X) = np$ and $V(X) = npq$ so $\sigma =$ 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 $np = (100)(\frac{1}{2}) = 50$. 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. Use web instead. Google "binomial distribution".

Example 3.32. Suppose that 20% of all copies of a particular textbook fail a certain binding strength text. Let X denote the number among 15 randomly selected copies that fail the test. Find

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

Solution. $X \sim \text{Bin}(15, .2)$. We want to compute $P(4 \le X \le 7)$ using the table on page 664. So how do we write $P(4 \le X \le 7)$ in terms of the form $P(X \le a)$.

Answer (#)

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

So

$$\begin{aligned} P(4 \leq X \leq 7) &= B(7; 15, .2) - B(3; 15, .2) \\ &= .996 - .648 \\ &= .348 \end{aligned}$$

N.B. **Understand (#).** This is the key to using computers and statistical calculators to compute.

2. The Hypergeometric Distribution.

Example

Millson to draw diagram

$$N = \text{chips}, \quad M = \text{red chips}, \quad L = \text{white chips}$$

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 $X = \#$ of red chips we get. Find the probability distriution of X.

Proposition.

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

if

$$\underbrace{\max(0, n - L) \leq k \leq \min(n, m)} \tag{$\flat$}$$

This means $k \leq$ both n and M and both 0 and $n - L \leq k$. These are the possible val;ues of k, that is, if k doesn't satisfy $\flat$ 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

$$\begin{aligned} \text{red chip} &\longleftrightarrow \text{heart} \\ \text{white chip} &\longleftrightarrow \text{ non-heart} \end{aligned}$$

So we use the principle of restricted choice

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

This agrees with $(*)$. But $(*)$ is harder because we have to consider the case where there are $k<n$ 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 chips $\binom{L}{n-k}$ ways. So

$$\#\left(\begin{array}{l}\text{choices of exactly } k \text{ red chips}\\ \text{in the } n \text{ chips}\end{array}\right)=\binom{M}{k}\binom{L}{n-k}.$$

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 hypergeometric distribution with parameters n, M, N. In the text the pmf is denoted*

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

What about the conditions

$$\max(0,n-L)\le k\le\min(n,m). \tag{$\flat$}$$

This really means

$$k\le \text{both } n \text{ and } M \tag{$\flat_1$}$$

and

$$\text{both } 0 \text{ and } n-L\le k. \tag{$\flat_2$}$$

$\flat_1$ says

$$\begin{aligned} k\le n &\longleftrightarrow && \text{we can't choose more than } n \text{ red chip}\\ & && \text{because we are only choosing } n \text{ chips in total}\\ k\le M &\longleftrightarrow && \text{because there are only } M \text{ red chips to choose from}\end{aligned}$$

and $\flat_2$

$k\ge 0$ 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 done yet, you have to choose $n-k$ white chips and there are only L white chips available so if $n-k>L$ we are sunk so we must have

$$n-k\le L\Longleftrightarrow k\ge n-L.$$

This is the second inequality of $(\flat_2)$. If it is satisfied, we can go ahead and choose the $n-k$ white chips to the inequalities in $(\flat)$ are necessary and sufficient.

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

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

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

If you put

$$p = \frac{M}{N} = \text{the probability of getting a red disk on the first draw}$$

then we may rewrite the above formulas as

$$\left.\begin{aligned} E(X) &= np \\ V(X) &= \left(\tfrac{N-n}{N-1}\right) npq \end{aligned}\right\} \text{reminiscent of the binomial distribution.}$$

Another Way to Derive (*)

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

Example . *Take $n = 2$*

$$\begin{aligned} P(X=0) &= \frac{L}{N}\frac{L-1}{N-1} \\ P(X=2) &= \frac{M}{N}\frac{M-1}{N-1} \\ P(X=1) &= P(RW) + P(WR) \\ &= \frac{M}{N}\frac{L}{N-1} + \frac{M}{N}\frac{L}{N-1} \\ &= 2\frac{M}{N}\frac{L}{N-1} \end{aligned}$$

In general, we clalim that all the words with $kR's$ and $n - kW'2$ have the same probability. Indeed each of these probabilities are fractions with the same denominator

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

and they have the same factors in the numerator scrambled up $M(M-1)(M-k+1)$ and $L(L-1)\cdots,(L-n-k+1)$. But the order of the factors doesn't matter so

$$\begin{aligned} P(X=k) &= \binom{n}{k} P(R\overset{k}{\cdots}R\, W\cdots W) \\ &= \binom{n}{k}\frac{M(M-1)\cdots(M-k+1)L(L-1)\cdots(L-n-k+1)}{N(N-1)\cdots N(-n+1)}. \end{aligned}$$

Why is (*) equal to this?

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

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

Exercise in fractions

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

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

Obviously, the first way (*) is easier so if you are doing a real-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.

An Important General Problem

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

This explain (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 chips have been removed because so few (relatively) have been removed."

In this case

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

(because N is huge $\frac{1}{N}$ and $\frac{N}{N} \approx 0$). So $V(X) \approx npq$. 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."