

STAT 515 Lec 5

Bernoulli trials, binomial and hypergeometric distributions

Karl Gregory

The Bernoulli trial



Figure 1: Jakob I. Bernoulli (1655 – 1705)

Definition: Bernoulli Trial

A Bernoulli trial is a statistical experiment for which there are two possible outcomes. We often give the outcomes the names “success” and “failure”. We denote by p the probability of a success.

Example. Shoot a freethrow and call a basket a success (LeBron has a higher p than you).

Example. Flip a coin and call heads a success. The probability of success is $p = 0.5$ if the coin is fair.

We often encode the outcome of a Bernoulli trial in a random variable which takes the value 1 if the outcome was a “success” and 0 if the outcome was a “failure”. The probability

distribution of such a random variable is called a *Bernoulli distribution*.

Definition: Bernoulli Distribution

Let X be a random variable taking the value 1 with probability p and the value 0 with probability $1 - p$. Then the probabilities $P(X = x)$ for x in $\mathcal{X} = \{0, 1\}$ are given by the expression

$$P(X = x) = p^x(1 - p)^{1-x}.$$

The probability distribution of X is called the *Bernoulli distribution* with probability of success equal to p .

If X is a random variable having the Bernoulli distribution with probability of success equal to p , then we write $X \sim \text{Bernoulli}(p)$.

The binomial distribution

We are often interested in the number of successes we observe in a sequence of independent Bernoulli trials.

Example. Flip a coin four times and let X be the number of heads, i.e. successes. We can write down the probability distribution of X by listing the possible sequences of successes and failures:

$$\mathcal{S} = \left\{ \begin{array}{l} TTTT, \\ TTTT, \quad TTHT, \quad THTT, \quad HTTT, \\ TTHH, \quad THTH, \quad THHT, \quad HTTH, \quad HTHT, \quad HHTT \\ HHHT, \quad HHTH, \quad HTHH, \quad THHH, \\ HHHH \end{array} \right\}.$$

Since each sequence of successes and failures is equally likely, we have the following:

x	0	1	2	3	4
$P(X = x)$	1/16	4/16	6/16	4/16	1/16
$P(X \leq x)$	1/16	5/16	11/16	15/16	16/16

Suppose now that we were to flip a coin 12 times and we wanted to know the probabilities of getting 0, 1, 2, or 3 heads, and so on. As the number of coin tosses increases, it quickly becomes very burdensome to list all the possible sequences of coin flips. In fact, flipping a coin 12 times is like drawing a ball from each of 12 urns, each of which contains two balls. There are $2^{12} = 4,096$ ways to do this. Nobody wants to list all of these possible sequences!

The binomial distribution comes to the rescue. It gives us the probability distribution of the number of successes in n independent Bernoulli trials when the probability of success is equal to p .

Definition: Binomial distribution

Let X be the number of successes in n independent Bernoulli trials having success probability p . Then the probabilities $P(X = x)$ for x in $\mathcal{X} = 0, 1, \dots, n$ are given by

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

The distribution with these probabilities is called the *binomial distribution* with the number of trials equal to n and with success probability equal to p .

If X is a random variable with the binomial distribution with the number of trials equal to n and with success probability equal to p , then we write $X \sim \text{Binomial}(n, p)$.

Let's dissect the expression

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

The quantity $\binom{n}{x}$ is the number of ways we can choose x of the n trials to be the successes. That is, it is the number of length- n success-failure sequences containing exactly x successes. Each sequence of successes and failures containing exactly x successes occurs with probability $p^x (1 - p)^{n-x}$, which is p raised to the number of successes times $(1 - p)$ raised to the number of failures. So the sum of the probabilities of all outcomes having x successes out of n trials is given by $\binom{n}{x} p^x (1 - p)^{n-x}$.

If X has the $\text{Binomial}(n, p)$ distribution, we can get compute the cumulative probabilities $P(X \leq x)$ for $x = 0, 1, \dots, n$ as

$$P(X \leq x) = \sum_{i=0}^x \binom{n}{i} p^i (1 - p)^{n-i}.$$

Exercise. Suppose you shoot 10 freethrows and you let X be the number of baskets you make. Suppose that you make freethrows with probability 0.3 and that all your shots are independent.

1. What is the probability that you will make 3 of the 10 baskets?

Answer: The 10 freethrows are 10 independent Bernoulli trials, so we can use the binomial distribution, calling a basket a success:

$$\begin{aligned}P(X = 3) &= \binom{10}{3}(0.3)^3(1 - 0.3)^{10-3} \\ &= \frac{10 \cdot 9 \cdot 8}{3 \cdot 2}(0.027)(0.0823543) \\ &= 0.267.\end{aligned}$$

We can tell R to compute this with the command `dbinom(x=3,size=10,prob=.3)`.

2. What is the probability that you make 6 or more baskets?

Answer:

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - \sum_{i=0}^5 \binom{10}{i} (0.3)^i (1 - 0.3)^{10-i} = 0.04734899.$$

The reason we change $P(X \geq 6)$ to $1 - P(X \leq 5)$ is that R has a function `pbinom()` to compute the cumulative probabilities of binomial random variables. We can compute the above with

$$1 - \text{pbinom}(q = 5, \text{size}=10, \text{prob}=.3)$$

Note that we needed to put `q = 5`.

Mean and variance of a binomial random variable

If $X \sim \text{Binomial}(n, p)$, then

$$\begin{aligned}\mu_X &= \mathbb{E}X = np \\ \sigma_X^2 &= \text{Var } X = np(1 - p).\end{aligned}$$

These expressions can be obtained by computing

$$\begin{aligned}\mathbb{E}X &= \sum_{x=0}^n x \binom{n}{x} p^x (1 - p)^{n-x} \\ \text{Var } X &= \sum_{x=0}^n (x - \mu_X)^2 \binom{n}{x} p^x (1 - p)^{n-x},\end{aligned}$$

but these are fairly tricky. Don't worry about it!

Random sampling and the hypergeometric distribution

Let us return now to one of the main goals of Statistics: getting information about a population from a random sample. Let us suppose that we have a population with N elements (people, critters, financial transactions, experimental units, etc.), and we draw a sample of size n . Suppose further that the members of the population may be classified as “successes” and “failures”, that is, suppose we are interested in some binary (taking two values) trait of the population members. Eventually, the goal will be to make a guess at the proportion of successes present in the population based on our sample of size n . To this end, we would like to understand how the number of successes in our sample will behave. That is, if X is the number of successes in our sample of size n drawn from a population of size N which contains, say M successes, we would like to know the probability distribution of X .

Exercise. Suppose we take a sample of size 5 from a population with 100 people of whom 5 are smokers. What is the probability that exactly 1 person in our sample will be a smoker?

Answer: There are $\binom{5}{1}$ ways to choose 1 of the 5 smokers. If we choose one smoker, then the other four in our sample will be nonsmokers, of which there are a total of 95 in the population. We can select 4 people from 95 people in $\binom{95}{4}$ ways. So there are $\binom{5}{1} \times \binom{95}{4}$ ways to get exactly one of the 5 smokers in a sample of 5. If divide this by the total number of ways in which we can sample 5 people from 100 people, we get

$$\begin{aligned} P(1 \text{ smoker in sample of } 5) &= \frac{\binom{5}{1} \binom{95}{4}}{\binom{100}{5}} = \frac{5! \cdot 95!}{1!4! \cdot 4!91!} = \frac{5 \cdot 95! \cdot 95! \cdot 5!}{4! \cdot 91! \cdot 100!} \\ &= \frac{5 \cdot 5 \cdot 95 \cdot 94 \cdot 93 \cdot 92}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} = 0.2114. \end{aligned}$$

More generally, we can write down the distribution of X as follows:

Definition: Hypergeometric distribution

Suppose we draw K marbles from a bag of N marbles, M of which are red, and let X be the number of red marbles drawn. Then the probabilities $P(X = x)$ are given by

$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} \quad \text{for } x = \max\{K - (N - M), 0\}, \dots, \min\{M, K\}.$$

The distribution with these probabilities is called the hypergeometric distribution.

The support \mathcal{X} of the hypergeometric distribution is actually kind of complicated; if the number of marbles K that we draw is greater than or equal to the number of non-red

marbles $N - M$ in the bag, then we must draw at least $K - (N - M)$ red marbles. Likewise, if the number of marbles K that we draw is less than the number of red marbles M in the bag, then we can draw at most K red marbles. So the possible values of X are given by

$$\max\{K - (N - M), 0\}, \dots, \min\{M, K\}.$$

If X is a random variable with a hypergeometric distribution corresponding to drawing K marbles from a bag of N , of which M are red, and letting X equal the number of red marbles drawn, then we write $X \sim \text{Hypergeometric}(N, M, K)$. Furthermore, the mean and variance of X are given by

$$\begin{aligned} \mathbb{E}X &= K \frac{M}{N} \\ \text{Var } X &= K \frac{M}{N} \left[\frac{(N - K)(N - M)}{N(N - 1)} \right]. \end{aligned}$$

Example. Suppose that in a class of 20 students, 12 of the students have 28 teeth. Consider drawing a sample of 5 students and letting X be the number of students among the 5 who have 28 teeth. We can use the hypergeometric distribution to tabulate the probability distribution of X :

x	0	1	2	3	4	5
$P(X = x)$	0.004	0.054	0.238	0.397	0.255	0.051
$P(X \leq x)$	0.004	0.058	0.296	0.693	0.949	1.000

We can get the values in the top row with the R commands

```
round(dhyper(x=0:5 , m = 12 , n = 20 - 12 , k = 5), 3)
```

or by putting `x <- 0:5` and then executing

```
round(choose(12, x)*choose(20-12, 5-x)/choose(20, 5), 3)
```

The values in the bottom row can be obtained with

```
round(phyper(q=0:5 , m = 12 , n = 20 - 12 , k = 5), 3)
```

or by putting `x <- 0:5` and then executing

```
round(cumsum(choose(12, x)*choose(20-12, 5-x)/choose(20, 5)), 3)
```

Note that in the `dhyper()` function, `m` is our M , `n` is our $N - M$, and `k` is our K . Type `?dhyper` to access the help file for the function `dhyper`, which describes the quantities in terms of drawing balls from urns. I told you statisticians like to draw balls from urns! We can also get, say, $P(X \leq 3)$, with the R command

```
phyper(q=3 , m = 12 , n = 20 - 12 , k = 5)
```

which gives $P(X \leq 3) = 0.6934985$.

Sampling without versus with replacement

The following exercise explores the difference between sampling with and without replacement from a large population.

Exercise. Suppose we are interested in estimating the proportion of USC undergraduates who are registered to vote. Let's say there are 24,830 USC undergraduates, and suppose that 9,932 of them are registered to vote (we don't actually know the true number). Suppose we sample the students sequentially and we let S_1 be the event that the first student drawn is registered to vote and let S_2 be the event that the second student drawn is registered to vote. Are the events S_1 and S_2 independent?

Answer: We have

$$P(S_1) = \frac{9,932}{24,830} = 0.40.$$

We have also

$$P(S_2|S_1) = \frac{9,931}{24,829} \quad \text{and} \quad P(S_2|S_1^c) = \frac{9,932}{24,829}.$$

We can already tell that S_1 and S_2 are not independent, because the probability of S_2 changes depending on whether S_1 has occurred or not. But let's work toward checking for independence using the test

$$P(S_1 \cap S_2) \stackrel{??}{=} P(S_1)P(S_2).$$

We can write:

$$\begin{aligned} P(S_2 \cap S_1) &= P(S_2|S_1)P(S_1) = \frac{9,931}{24,829} \times \frac{9,932}{24,830} = 0.1599903 \\ P(S_2 \cap S_1^c) &= P(S_2|S_1^c)P(S_1^c) = \frac{9,932}{24,829} \times \left(1 - \frac{9,932}{24,830}\right) = 0.2400097 \end{aligned}$$

Now since

$$P(S_2) = P(S_2 \cap S_1) + P(S_2 \cap S_1^c),$$

we have

$$P(S_2) = 0.1599903 + 0.2400097 = 0.40.$$

So

$$P(S_1)P(S_2) = 0.40 \times 0.40 = .16 \quad \text{and} \quad P(S_1 \cap S_2) = 0.1599903.$$

The answer is that S_1 and S_2 are *almost* independent.

Imagine a sampling scheme in which we draw a student, write down whether they are registered to vote or not, and then place them back in the population. We repeat this n times. This is called *sampling with replacement*. Under sampling with replacement, each draw can be regarded as a Bernoulli trial, and all the draws/trials are independent. Under *sampling*

without replacement, which is what we actually do in practice, each draw can still be regarded as a Bernoulli trial, but our entire sample represents a sequence of dependent Bernoulli trials. However, if the population has a large number of successes and failures in it, the draws under *sampling without replacement* can be regarded as independent Bernoulli trials, even though this is not truly the case.

Example. Sample 6 USC undergraduates *without replacement* and let X be the number of students in the sample who are registered to vote. If in fact 9,932 of the 24,830 USC undergraduates are registered to vote, then X follows a hypergeometric distribution with $N = 24,830$, $M = 9,932$, and $K = 6$, giving

x	0	1	2	3	4	5	6
$P(X = x)$	0.0466	0.1866	0.3111	0.2765	0.1382	0.0368	0.0041
$P(X \leq x)$	0.0466	0.2332	0.5443	0.8208	0.9591	0.9959	1.0000

which can be obtained using the R commands

```
round(dhyper(x=0:6, m = 9932, n = 24830 - 9932, k = 6), 4)
```

and

```
round(phyper(q=0:6, m = 9932, n = 24830 - 9932, k = 6), 4)
```

Now sample 6 USC undergraduates *with replacement* and let X be the number of students in the sample who are registered to vote. If in fact 9,932 of the 24,830 USC undergraduates are registered to vote, then X follows a binomial distribution with $n = 6$ and $p = 0.40$, giving

x	0	1	2	3	4	5	6
$P(X = x)$	0.0467	0.1866	0.3110	0.2765	0.1382	0.0369	0.0041
$P(X \leq x)$	0.0467	0.2333	0.5443	0.8208	0.9590	0.9959	1.0000

which can be obtained using the R commands

```
round(dbinom(x=0:6, size=6, prob=.40), 4)
```

and

```
round(pbinom(q=0:6, size=6, prob=.40), 4)
```

We see that the numbers are almost exactly the same.

Treating draws without replacement as independent Bernoulli trials

When the population contains a large number of successes and failures, we can regard draws taken *without replacement* as independent Bernoulli trials, even though they are not in fact independent: We get away with this because they are very close to being independent. Therefore, we will use the binomial distribution much much more than the hypergeometric distribution in this class.

Probability mass functions of discrete random variables

In these notes we have considered the Bernoulli, Binomial, and Hypergeometric distributions. For each of these distributions, a different function was used to compute the probabilities $P(X = x)$ for x in \mathcal{X} . We refer to such functions in general as probability mass functions:

Definition: Probability mass function

The probability mass function (pmf) of a discrete random variable X with support \mathcal{X} is the function p which satisfies

$$P(X = x) = p(x) \quad \text{for all } x \in \mathcal{X}.$$

Example. If $X \sim \text{Binomial}(n, p)$, then the probability mass function of X is

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

for x in \mathcal{X} .