The Hypergeometric Distribution

Drawing Balls from an Urn

The classical example of a sequence of non-independent trials involves drawing colored balls from a box, traditionally called an ***urn***.

Suppose that an urn contains a total of N balls, of which r are red balls and $N - r$ are black balls. We draw a sample of n balls from the urn. We ***do not replace*** each ball after drawing it.

The random variable X describes the number of red balls in our sample.

If we were to replace each ball after drawing, X would follow a binomial distribution. But now the probability of drawing a red ball depends on the previous outcomes.

Drawing Balls from an Urn

Given the number of objects N, the sample size n and the number r of red balls, we can apply Cardano's principle to calculate $P[X = x]$.

We will assume that

$$r > n \qquad \text{and} \qquad N - r > n,$$

so that we could have 0 to n black or red balls in our sample.

Then

$$\begin{aligned} &P[\text{exactly } x \text{ red balls out of } n \text{ selected}] \\ &= \frac{(\# \text{ ways to select } x \text{ out of } r \text{ balls}) \cdot (\# \text{ ways to select } n - x \text{ out of } N - r \text{ balls})}{\# \text{ ways to select } n \text{ out of } N \text{ balls}} \\ &= \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \end{aligned}$$

The Hypergeometric Distribution

9.1. Definition. Let $N, n, r \in \mathbb{N} \setminus \{0\}$, $r, n \leq N$, and $n < \min\{r, N - r\}$.

A random variable (X, f_X) with

$$X : S \to \Omega = \{0, \dots, n\}$$

and density function $f_X : \Omega \to \mathbb{R}$ given by

$$f_X(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \tag{9.1}$$

is said to have a hypergeometric distribution with parameters N, n and r.

The Hypergeometric Identity

The hypergeometric distribution takes its name from the ***hypergeometric identity***:

$$\binom{a+b}{r} = \sum_{k=0}^{r} \binom{a}{k}\binom{b}{r-k} = \sum_{i+j=r} \binom{a}{i}\binom{b}{j}. \tag{9.2}$$

To understand this identity, note that

$$\begin{aligned}\sum_{r=0}^{a+b} \binom{a+b}{r} x^r &= (1+x)^{a+b} = (1+x)^a(1+x)^b \\ &= \Big(\sum_{i=0}^{a} \binom{a}{i} x^i\Big)\Big(\sum_{j=0}^{b} \binom{b}{j} x^j\Big)\end{aligned}$$

The Hypergeometric Identity

Using the definition of the binomial coefficients (1.3), we see that $\binom{x}{i} = 0$ when $i > x$, so we may write

$$\begin{aligned}\sum_{r=0}^{\infty} \binom{a+b}{r} x^r &= \Big(\sum_{i=0}^{\infty} \binom{a}{i} x^i\Big)\Big(\sum_{j=0}^{\infty} \binom{b}{j} x^j\Big) \\ &= \sum_{r=0}^{\infty} \sum_{i+j=r} \binom{a}{i}\binom{b}{j} x^r,\end{aligned}$$

where we have used the Cauchy product of infinite series. Comparing term-by-term, (9.2) follows.

The hypergeometric identity is the main ingredient in showing that (9.1) actually defines a density function.

Non-independent Bernoulli Trials

Let us write

$$X = X_1 + X_2 + \cdots + X_n,$$

where each X_k is a Bernoulli random variable representing a single draw. Here "success" means drawing a red ball, yielding $X_k = 1$. If we draw a black ball on the kth draw, then $X_k = 0$.

We denote the probability of success by

$$p_k = P[X_k = 1]$$

Of course, the X_k are not independent - the result of each draw (X_k) influences the subsequent draws.

We therefore have to discuss the random vector $(X_1, X_2, \dots, X_n)$.

The Bernoulli Trials are Identical

To understand the distribution of the random vector $(X_1, \dots, X_n)$, we consider the sample space S: Suppose that we order the N balls in all conceivable ways, obtaining $N!$ permutations.

Effectively, we are drawing ***all balls*** out of the urn to obtain the sample space S, but only consider the events that include the ***first n*** balls to calculate the probabilities of our random vector.

The probability that $X_k = x$, where $x = 0$ or 1, is then given by the number of elements in the sample space that have x in the kth position.

Since the sample space consists of ***all*** possible permutations of the N objects, we see that this probability does not depend on k. Therefore,

$$p_k = p_1 = \frac{r}{N}.$$

This shows that the Bernoulli trials are identical.

Expectation and Variance

We can calculate

$$\mathrm{E}[X_k] = 0 \cdot (1 - p_k) + 1 \cdot p_k = p_k = p_1 = \frac{r}{N}$$

so

$$\mathrm{E}[X] = \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = n\frac{r}{N}.$$

It is interesting to note that this expectation is the same as it would be if we were replacing the balls after drawing, i.e., if the number of red balls were determined by the binomial distribution.

In order to calculate the variance, we first generalize (8.1) to

$$\begin{aligned}\operatorname{Var} X &= \operatorname{Var}(X_1 + \cdots + X_n) \\ &= \operatorname{Var} X_1 + \cdots + \operatorname{Var} X_n + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

Variance and Covariance

We need to calculate

$$\operatorname{Cov}(X_i, X_j) = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j].$$

For this, we note that $X_i X_j$ is also a Bernoulli variable, since

$$X_i X_j = \begin{cases} 1 & \text{if } X_i = 1 \text{ and } X_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\mathrm{E}[X_i X_j] = p_{ij} := P[X_i = 1 \text{ and } X_j = 1].$$

Variance and Covariance

As in the previous argument, the probability that $X_i = x$ and $X_j = y$ for $i \neq j$ and $x, y \in \{0, 1\}$ is given by the number of permutations among all $N!$ elements of the sample space that have x in the ith and y in the jth position. Again it is clear that this number is independent of i and j, so

$$p_{ij} = p_{12} = P[X_1 = 1 \text{ and } X_2 = 1] = \frac{r}{N} \cdot \frac{r-1}{N-1}.$$

Note that for $i = j$ we have

$$p_{ii} = p_{11} = P[X_1 = 1 \text{ and } X_1 = 1] = \frac{r}{N}.$$

Hence,

$$\operatorname{Var} X_i = \frac{r}{N}\left(1 - \frac{r}{N}\right), \qquad \operatorname{Cov}(X_i, X_j) = -\frac{1}{N} \cdot \frac{r(N-r)}{N(N-1)}.$$

Approximating the Hypergeometric Distribution

Since there are $\binom{n}{2} = n(n-1)/2$ pairs (i,j) with $i < j$, an easy calculation (do it yourself!) now gives

$$\operatorname{Var} X = n\frac{r}{N}\frac{N-r}{N}\frac{N-n}{N-1}$$

This expression is similar to that for the binomial distribution; if we were replacing the balls after drawing, we would have

$$p = \frac{r}{N}, \qquad q = \frac{N-r}{N}$$

and since the variance of the binomial distribution is npq, we see that the expression above differs by

$$\frac{N-n}{N-1}.$$

In fact, the binomial distribution may be used to approximate the hypergeometric distribution if the ***sampling fraction*** n/N is small (less than 0.05).

Approximating the Hypergeometric Distribution

9.2. Example. A production lot of 200 units has 8 defectives. A random sample of 10 units is selected, and we want to find the probability that the random sample will contain exactly one defective.

The true probability is

$$P[X = 1] = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = \frac{\binom{8}{1}\binom{192}{9}}{\binom{200}{10}} = 0.288.$$

We note that the sampling fraction is $n/N = 10/200 = 0.05$, so we can use the binomial approximation.

Then $p = r/N = 8/200 = 0.04$ and

$$P[X = 1] \approx \binom{10}{1}(0.04)^1(0.96)^9 = 0.277.$$