The Hypergeometric distribution

37.4

Introduction

The hypergeometric distribution enables us to deal with situations arising when we sample from batches with a known number of defective items. You will find that, in essence, the number of defective items in a batch is not a random variable - it is a known, fixed, number.

Prerequisites

Before starting this Section you should . . .

① understand the concepts of probability.

② understand the notation ${}^nC_r$ used in probability calculations.

Learning Outcomes

After completing this Section you should be able to . . .

✓ apply the hypergeometric distribution to simple examples.

1. The Hypergeometric Distribution

Suppose we are sampling without replacement from a batch of items containing a *variable number* of defectives. We are essentially assuming that we know the probability p that a given item is defective but not the *actual number* of defective items contained in the batch. The number of defective items in the batch is a random variable in this case.
When we sample from the batch, we are left with:

1. a smaller batch;
2. a (possibly) smaller (but still *variable*) number of defective items. The number of defective items is still a random variable.

While the probability of finding a given number of defectives in a sample drawn from the second batch will, (in general) be different from the probability of finding a given number of defectives in a sample drawn from the first batch, sampling from both batches may be described by the Binomial distribution for which:

$$P(X=r) = {}^{n}C_r(1-p)^{n-r}p^r$$

Sampling in this case varies the values of n and p in general but not the underlying distribution describing the sampling process.

Example A batch of 100 piston rings is known to contain 10 defective rings. If two piston rings are drawn from the batch, write down the probabilities that:

1. the first ring is defective;
2. the second ring is defective given that the first one is defective.

Solution

1. The probability that the first ring is defective is clearly $\dfrac{10}{100} = \dfrac{1}{10}$.
2. Assuming that the first ring selected is defective and we do not replace it, the probability that the second ring is defective is equally clearly $\dfrac{9}{99} = \dfrac{1}{11}$.

The Hypergeometric distribution may be thought of as arising from sampling from a batch of items where the number of defective items contained in the batch is known.

Essentially the number of defectives contained in the batch is not a random variable, it is fixed.

The calculations involved when using the Hypergeometric distribution are usually more complex than their Binomial counterparts.
If we sample without replacement we may proceed in general as follows:

- we may select n items from a population of N items in ${}^{N}C_n$ ways;
- we may select r defective items from M defective items in ${}^{M}C_r$ ways;
- we may select $n-r$ non-defective items from $N-M$ non-defective items in ${}^{N-M}C_{n-r}$ ways;

- hence we may select n items containing r defectives in ${}^{M}C_r \times {}^{N-M}C_{n-r}$ ways.
- hence the probability that we select a sample of size n containing r defective items from a population of N items known to contain M defective items is

$$\frac{{}^{M}C_r \times {}^{N-M}C_{n-r}}{{}^{N}C_n}$$

Key Point

The distribution given by

$$P(X = r) = \frac{{}^{M}C_r \times {}^{N-M}C_{n-r}}{{}^{N}C_n}$$

which describes the probability of obtaining a sample of size n containing r defective items from a population of size N known to contain M defective items is known as the **Hypergeometric distribution**.

Example A batch of 10 rocker cover gaskets contains 4 defective gaskets. If we draw samples of size 3 without replacement, from the batch of 10, find the probability that a sample contains 2 defective gaskets.

Solution

Using $P(X = r) = \dfrac{{}^{M}C_r \times {}^{N-M}C_{n-r}}{{}^{N}C_n}$ we know that $N = 10$, $M = 4$, $n = 3$ and $r = 2$.

Hence $P(X = 2) = \dfrac{{}^{4}C_2 \times {}^{6}C_1}{{}^{10}C_3} = \dfrac{6 \times 6}{120} = 0.3$

It is possible to derive formulae for the mean and variance of the Hypergeometric distribution. However, the calculations are more difficult than their Binomial counterparts, so we will simple state the results.

Key Point

The expectation (mean) and variance of the Hypergeometric random variable

$$P(X=r)=\frac{{}^{M}C_{r}\ \times\ {}^{N-M}C_{n-r}}{{}^{N}C_{n}}$$

are given by

$$E(X)=\mu=np \qquad \text{and} \qquad V(X)=np(1-p)\frac{N-M}{N-1} \qquad \text{where } \ p=\frac{M}{N}$$

Example For the previous example, concerning rocker cover gaskets find the expectation and variance of samples containing 2 defective gaskets.

Solution

Using $\quad P(X=r)=\dfrac{{}^{M}C_{r}\ \times\ {}^{N-M}C_{n-r}}{{}^{N}C_{n}}\quad$ we know that $N=10,\ M=4,\ n=3$ and $r=2$.

Hence

$$E(X)=np=3\times\frac{4}{10}=1.2$$

and

$$V(X)=np(1-p)\frac{N-M}{N-1}=3\times\frac{4}{10}\times\frac{6}{10}\times\frac{10-4}{10-1}=0.48$$

In the manufacture of car tyres, a particular production process is know to yield 10 tyres with defective walls in every batch of 100 tyres produced. From a production batch of 100 tyres, a sample of 4 is selected for testing to destruction. Find:

(a) the probability that the sample contains 1 defective tyre

(b) the expectation of the number of defectives in samples of size 4

(c) the variance of the number of defectives in samples of size 4.

Your solution

Sampling is clearly without replacement and we use the Hypergeometric distribution with $N = 100, M = 10, n = 4, r = 1$ and $p = 0.1$. Hence:

(a) $$P(X = r) = \frac{{}_{M}C_r \times {}_{N-M}C_{n-r}}{{}_{N}C_n} \quad \text{gives}$$

$$P(X = 1) = \frac{{}_{10}C_1 \times {}_{100-10}C_{4-1}}{{}_{100}C_4} = \frac{10 \times 117480}{3921225} \approx 0.3$$

(b) The expectation is $E(X) = np = 4 \times 0.1 = 0.4$

(c) The variance is $V(X) = np(1-p)\dfrac{N-M}{N-1} = 0.4 \times 0.9 \times \dfrac{90}{99} \approx 0.33$

A company (the producer) supplies microprocessors to a manufacturer (the consumer) of electronic equipment. The microprocessors are supplied in batches of 50. The consumer regards a batch as acceptable provided that there are not more than 5 defective microprocessors in the batch. Rather than test all of the microprocessors in the batch, 10 are selected at random and tested.

Find the probability that out of a sample of 10, $d = 0, 1, 2, 3, 4, 5$ are defective when there are actually 5 defective microprocessors in the batch.

Suppose that the consumer will accept the batch provided that not more than m defectives are found in the sample of 10.

(a) Find the probability that the batch is accepted when there are 5 defectives in the batch.

(b) Find the probability that the batch is rejected when there are 3 defectives in the batch.

Your solution

Let X = the numbers of defectives in a sample. Then

$$P(X=d) = \frac{^{45}C_{10-d} \times {}^{5}C_d}{^{50}C_{10}}$$

Hence

$$P(X=0) = \frac{^{45}C_{10} \times {}^{5}C_0}{^{50}C_{10}} = 0.311 \qquad P(X=1) = \frac{^{45}C_{9} \times {}^{5}C_1}{^{50}C_{10}} = 0.431$$

$$P(X=2) = \frac{^{45}C_{8} \times {}^{5}C_2}{^{50}C_{10}} = 0.210 \qquad P(X=3) = \frac{^{45}C_{7} \times {}^{5}C_3}{^{50}C_{10}} = 0.044$$

$$P(X=4) = \frac{^{45}C_{6} \times {}^{5}C_4}{^{50}C_{10}} = 0.004 \qquad P(X=5) = \frac{^{45}C_{5} \times {}^{5}C_5}{^{50}C_{10}} = 0.0001$$

Case $D = 5$

Prob(Accept batch with 5 defectives) is

$$\sum_{d=0}^{m} P(X=d) = \sum_{d=0}^{m} \frac{^{45}C_{10-d} \times {}^{5}C_d}{^{50}C_{10}} \qquad m \leq 5$$

Case $D = 3$

Prob(Reject batch with 3 defectives) is

$$1 - \sum_{d=0}^{m} P(X=d) = 1 - \sum_{d=0}^{m} \frac{^{47}C_{10-d} \times {}^{3}C_d}{^{50}C_{10}} \qquad m \leq 3$$

Exercises

1. A company buys batches of n components. Before a batch is accepted, m of the components are selected at random from the batch and tested. The batch is rejected if more than d components in the sample are found to be below standard.

   (a) Find the probability that a batch which actually contains six below-standard components is rejected when $n = 20$, $m = 5$ and $d = 1$.

   (b) Find the probability that a batch which actually contains nine below-standard components is rejected when $n = 30$, $m = 10$ and $d = 1$.

Answers

1. (a) Let the number of below-standard components in the sample be X. The probability of acceptance is

$$\begin{aligned}
\Pr(X=0)+\Pr(X=1) &= \frac{\binom{14}{5}\binom{6}{0}}{\binom{20}{5}} + \frac{\binom{14}{4}\binom{6}{1}}{\binom{20}{5}} \\
&= \frac{\frac{14}{5}\times\frac{13}{4}\times\frac{12}{3}\times\frac{11}{2}\times\frac{10}{1}+\frac{14}{4}\times\frac{13}{3}\times\frac{12}{2}\times\frac{12}{2}\times\frac{11}{1}\times\frac{6}{1}}{\frac{20}{5}\times\frac{19}{4}\times\frac{18}{3}\times\frac{17}{2}\times\frac{16}{1}} \\
&= \frac{2002+6006}{15504} \\
&= 0.5165
\end{aligned}$$

Hence the probability of rejection is $1-0.5165=0.4835$.

(b) Let the number of below-standard components in the sample be X. The probability of acceptance is

$$\Pr(X=0)+\Pr(X=1) = \frac{\binom{21}{10}\binom{9}{0}}{\binom{30}{10}} + \frac{\binom{21}{9}\binom{9}{1}}{\binom{30}{10}}$$

Now

$$\begin{aligned}
\binom{21}{10}\binom{9}{0} &= \frac{21}{10}\times\frac{20}{9}\times\frac{19}{8}\times\frac{18}{7}\times\frac{17}{6}\times\frac{16}{5}\times\frac{15}{4}\times\frac{14}{3}\times\frac{13}{2}\times\frac{12}{1} \\
&= 352716 \\
\binom{21}{9}\binom{9}{1} &= \frac{21}{9}\times\frac{20}{8}\times\frac{19}{7}\times\frac{18}{6}\times\frac{17}{5}\times\frac{16}{4}\times\frac{15}{3}\times\frac{14}{2}\times\frac{13}{1}\times\frac{9}{1} \\
&= 2645370 \\
\binom{30}{10} &= \frac{30}{10}\times\frac{29}{9}\times\frac{28}{8}\times\frac{27}{7}\times\frac{26}{6}\times\frac{25}{5}\times\frac{24}{4}\times\frac{23}{3}\times\frac{22}{2}\times\frac{21}{1}
\end{aligned}$$

So the probability of acceptance is

$$\frac{352716+2645370}{30045015}=0.0998.$$

Hence the probability of rejection is $1-0.0998=0.9002$.