Contents 37

Discrete Probability Distributions

Learning outcomes

In this Workbook you will learn what a discrete random variable is. You will find how to calculate the expectation and variance of a discrete random variable. You will then examine two of the most important examples of discrete random variables: the binomial distribution and Poisson distribution.

The Poisson distribution can be deduced from the binomial distribution and is often used as a way of finding good approximations to the binomial probabilities. The binomial is a finite discrete random variable whereas the Poisson distribution has an infinite number of possibilities.

Finally you will learn about anotherimportant distribution - the hypergeometric.

Discrete Probability Distributions

37.1

Introduction

It is often possible to model real systems by using the same or similar random experiments and their associated random variables. Numerical random variables may be classified in two broad but distinct categories called discrete random variables and continuous random variables. Often, discrete random variables are associated with counting while continuous random variables are associated with measuring. In HELM 42. you will meet contingency tables and deal with non-numerical random variables. Generally speaking, discrete random variables can take values which are separate and can be listed. Strictly speaking, the real situation is a little more complex but it is sufficient for our purposes to equate the word discrete with a finite list. In contrast, continuous random variables can take values anywhere within a specified range. This Section will familiarize you with the idea of a discrete random variable and the associated probability distributions. The Workbook makes no attempt to cover the whole of this large and important branch of statistics but concentrates on the discrete distributions most commonly met in engineering. These are the binomial, Poisson and hypergeometric distributions.

Prerequisites

Before starting this Section you should ...

- understand the concepts of probability

Learning Outcomes

On completion you should be able to ...

- explain what is meant by the term discrete random variable
- explain what is meant by the term discrete probability distribution
- use some of the discrete probability distributions which are important to engineers

1. Discrete probability distributions

We shall look at discrete distributions in this Workbook and continuous distributions in HELM 38. In order to get a good understanding of discrete distributions it is advisable to familiarise yourself with two related topics: permutations and combinations. Essentially we shall be using this area of mathematics as a calculating device which will enable us to deal sensibly with situations where *choice* leads to the use of very large numbers of possibilities. We shall use combinations to express and manipulate these numbers in a compact and efficient way.

Permutations and Combinations

You may recall from HELM 35.2 concerned with probability that if we define the probability that an event A occurs by using the definition:

$$\mathsf{P}(A) = \frac{\text{The number of equally likely experimental outcomes favourable to } A}{\text{The total number of equally likely outcomes forming the sample space}} = \frac{a}{n}$$

then we can only find $\mathsf{P}(A)$ provided that we can find both a and n. In practice, these numbers can be very large and difficult if not impossible to find by a simple counting process. Permutations and combinations help us to calculate probabilities in cases where counting is simply not a realistic possibility.

Before discussing permutations, we will look briefly at the idea and notation of a factorial.

Factorials

The **factorial** of an integer n commonly called 'factorial n' and written $n!$ is defined as follows:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1 \qquad n \geq 1$$

Simple examples are:

$$3! = 3\times2\times1 = 24 \qquad 5! = 5\times4\times3\times2\times1 = 120 \qquad 8! = 8\times7\times6\times5\times4\times3\times2\times1 = 40320$$

As you can see, factorial notation enables us to express large numbers in a very compact format. You will see that this characteristic is very useful when we discuss the topic of permutations. A further point is that the definition above falls down when $n = 0$ and we define

$$0! = 1$$

Permutations

A **permutation** of a set of distinct objects places the objects **in order**. For example the set of three numbers $\{1, 2, 3\}$ can be placed in the following orders:

1,2,3   1,3,2   2,1,3   2,3,1   3,2,1   3,1,2

Note that we can choose the first item in 3 ways, the second in 2 ways and the third in 1 way. This gives us $3\times2\times1 = 3! = 6$ distinct orders. We say that the set $\{1, 2, 3\}$ has the distinct permutations

1,2,3   1,3,2   2,1,3   2,3,1   3,2,1   3,1,2

Example 1

Write out the possible permutations of the letters A, B, C and D.

Solution

The possible permutations are

$$\begin{array}{llllll} ABCD & ABDC & ADBC & ADCB & ACBD & ACDB \\ BADC & BACD & BCDA & BCAD & BDAC & BDCA \\ CABD & CADB & CDBA & CDAB & CBAD & CBDA \\ DABC & DACB & DCAB & DCBA & DBAC & DBCA \end{array}$$

There are $4! = 24$ permutations of the four letters A, B, C and D.

In general we can order n distinct objects in $n!$ ways.
Suppose we have r different types of object. It follows that if we have n_1 objects of one kind, n_2 of another kind and so on then the n_1 objects can be ordered in $n_1!$ ways, the n_2 objects in $n_2!$ ways and so on. If $n_1 + n_2 + \cdots + n_r = n$ and if p is the number of permutations possible from n objects we may write

$$p \times (n_1! \times n_2! \times \cdots \times n_r!) = n!$$

and so p is given by the formula

$$p = \frac{n!}{n_1! \times n_2! \times \cdots \times n_r!}$$

Very often we will find it useful to be able to calculate the number of permutations of n objects taken r at a time. Assuming that we do not allow repetitions, we may choose the first object in n ways, the second in $n-1$ ways, the third in $n-2$ ways and so on so that the r^{th} object may be chosen in $n - r + 1$ ways.

Example 2

Find the number of permutations of the four letters A, B, C and D taken three at a time.

Solution

We may choose the first letter in 4 ways, either A, B, C or D. Suppose, for the purposes of illustration we choose A. We may choose the second letter in 3 ways, either B, C or D. Suppose, for the purposes of illustration we choose B. We may choose the third letter in 2 ways, either C or D. Suppose, for the purposes of illustration we choose C. The total number of choices made is $4 \times 3 \times 2 = 24$.

In general the numbers of permutations of n objects taken r at a time is

$$n(n-1)(n-2)\dots(n-r+1) \qquad \text{which is the same as} \quad \frac{n!}{(n-r)!}$$

This is usually denoted by ${}^nP_r$ so that

$${}^nP_r = \frac{n!}{(n-r)!}$$

If we allow repetitions the number of permutations becomes n^r (can you see why?).

Example 3

Find the number of permutations of the four letters A, B, C and D taken two at a time.

Solution

We may choose the first letter in 4 ways and the second letter in 3 ways giving us

$$4 \times 3 = \frac{4 \times 3 \times 2 \times 1}{1 \times 2} = \frac{4!}{2!} = 12 \quad \text{permutations}$$

Combinations

A **combination** of objects takes **no account of order** whereas a permutation does. The formula ${}^nP_r = \dfrac{n!}{(n-r)!}$ gives us the number of ordered sets of r objects chosen from n. Suppose the number of sets of r objects (taken from n objects) in which order is not taken into account is C. It follows that

$$C \times r! = \frac{n!}{(n-r)!} \quad \text{and so } C \text{ is given by the formula} \qquad C = \frac{n!}{r!(n-r)!}$$

We normally denote the right-hand side of this expression by ${}^nC_r$ so that

$${}^nC_r = \frac{n!}{r!(n-r)!} \qquad \text{A common alternative notation for } {}^nC_r \text{ is } \binom{n}{r}.$$

Example 4

How many car registrations are there beginning with $NP05$ followed by three letters? Note that, conventionally, I, O and Q may not be chosen.

Solution

We have to choose 3 letters from 23 allowing repetition. Hence the number of registrations beginning with $NP05$ must be $23^3 = 12167$.

Task

(a) How many different signals consisting of five symbols can be sent using the dot and dash of Morse code?

(b) How many can be sent if five symbols *or less* can be sent?

Your solution

Answer

(a) Clearly, the order of the symbols is important. We can choose each symbol in two ways, either a dot or a dash. The number of distinct signals is

$2 \times 2 \times 2 \times 2 \times = 2^5 = 32$

(b) If five *or less* symbols may be used, the total number of signals may be calculated as follows:

Using one symbol:	2 ways
Using two symbols:	$2 \times 2 = 4$ ways
Using three symbols:	$2 \times 2 \times 2 = 8$ ways
Using four symbols:	$2 \times 2 \times 2 \times 2 = 16$ ways
Using five symbols:	$2 \times 2 \times 2 \times 2 \times 2 = 32$ ways

The total number of signals which may be sent is 62.

Task

A box contains 50 resistors of which 20 are deemed to be 'very high quality' , 20 'high quality' and 10 'standard'. In how many ways can a batch of 5 resistors be chosen if it is to contain 2 'very high quality', 2 'high quality' and 1 'standard' resistor?

Your solution

Answers The order in which the resistors are chosen does not matter so that the number of ways in which the batch of 5 can be chosen is:

$$^{20}C_2 \times ^{20}C_2 \times ^{10}C_1 = \frac{20!}{18! \times 2!} \times \frac{20!}{18! \times 2!} \times \frac{10!}{9! \times 1!} = \frac{20 \times 19}{1 \times 2} \times \frac{20 \times 19}{1 \times 2} \times \frac{10}{1} = 361000$$

2. Random variables

A random variable X is a quantity whose value cannot be predicted with certainty. We assume that for every real number a the probability P$(X = a)$ in a trial is well-defined. In practice, engineers are often concerned with two broad types of variables and their probability distributions: discrete random variables and their distributions, and continuous random variables and their distributions. Discrete distributions arise from experiments involving counting, for example, road deaths, car production and aircraft sales, while continuous distributions arise from experiments involving measurement, for example, voltage, corrosion and oil pressure.

Discrete random variables and probability distributions

A random variable X and its distribution are said to be discrete if the values of X can be presented as an ordered list say $x_1, x_2, x_3, \ldots$ with probability values $p_1, p_2, p_3, \ldots$. That is P$(X = x_i) = p_i$. For example, the number of times a particular machine fails during the course of one calendar year is a discrete random variable.

More generally a discrete distribution $f(x)$ may be defined by:

$$f(x) = \begin{cases} p_i & \text{if } x = x_i \quad i = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

The distribution function $F(x)$ (sometimes called the cumulative distribution function) is obtained by taking sums as defined by

$$F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} p_i$$

We sum the probabilities p_i for which x_i is less than or equal to x. This gives a step function with jumps of size p_i at each value x_i of X. The step function is defined for all values, not just the values x_i of X.

Key Point 1

Probability Distribution of a Discrete Random Variable

Let X be a random variable associated with an experiment. Let the values of X be denoted by $x_1, x_2, \ldots, x_n$ and let P$(X = x_i)$ be the probability that x_i occurs. We have two necessary conditions for a valid probability distribution:

- P$(X = x_i) \geq 0$ for all x_i
- $\displaystyle\sum_{i=1}^{n}$ P$(X = x_i) = 1$

Note that n may be uncountably large (infinite).

(These two statements are sufficient to guarantee that P$(X = x_i) \leq 1$ for all x_i.)

Example 5

Turbo Generators plc manufacture seven large turbines for a customer. Three of these turbines do not meet the customer's specification. Quality control inspectors choose two turbines at random. Let the discrete random variable X be defined to be the number of turbines inspected which meet the customer's specification.

(a) Find the probabilities that X takes the values $0, 1$ or 2.

(b) Find and graph the cumulative distribution function.

Solution

(a) The possible values of X are clearly $0, 1$ or 2 and may occur as follows:

Sample Space	Value of X
Turbine faulty, Turbine faulty	0
Turbine faulty, Turbine good	1
Turbine good, Turbine faulty	1
Turbine good, Turbine good	2

We can easily calculate the probability that X takes the values $0, 1$ or 2 as follows:

$$\mathsf{P}(X = 0) = \frac{3}{7} \times \frac{2}{6} = \frac{1}{7} \qquad \mathsf{P}(X = 1) = \frac{4}{7} \times \frac{3}{6} + \frac{3}{7} \times \frac{4}{6} = \frac{4}{7} \qquad \mathsf{P}(X = 2) = \frac{4}{7} \times \frac{3}{6} = \frac{2}{7}$$

The values of $F(x) = \sum_{x_i \leq x} \mathsf{P}(X = x_i)$ are clearly

$$F(0) = \frac{1}{7} \qquad F(1) = \frac{5}{7} \qquad \text{and} \qquad F(2) = \frac{7}{7} = 1$$

(b) The graph of the step function $F(x)$ is shown below.

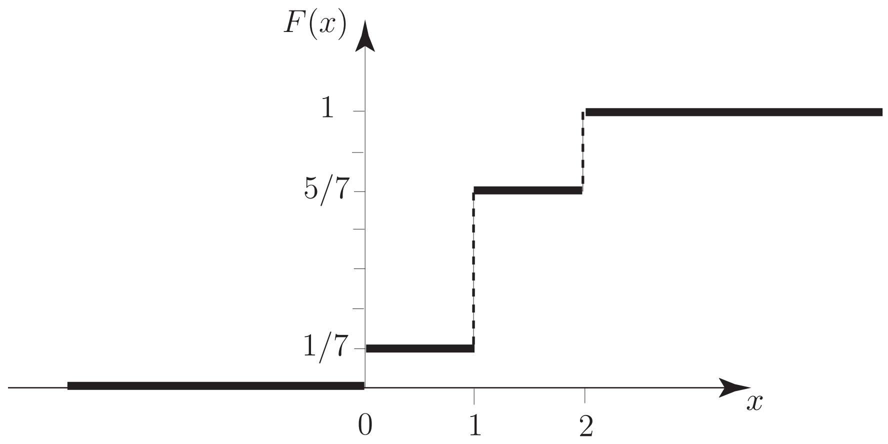

Figure 1

3. Mean and variance of a discrete probability distribution

If an experiment is performed N times in which the n possible outcomes $X = x_1, x_2, x_3, \ldots, x_n$ are observed with frequencies $f_1, f_2, f_3, \ldots, f_n$ respectively, we know that the mean of the distribution of outcomes is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i = \sum_{i=1}^{n}\left(\frac{f_i}{N}\right)x_i$$

(Note that $\sum_{i=1}^{n} f_i = f_1 + f_2 + \cdots + f_n = N$.)

The quantity $\frac{f_i}{N}$ is called the **relative frequency** of the observation x_i. Relative frequencies may be thought of as akin to probabilities; informally we would say that the chance of observing the outcome x_i is $\frac{f_i}{N}$. Formally, we consider what happens as the number of experiments becomes very large. In order to give meaning to the quantity $\frac{f_i}{N}$ we consider the limit (if it exists) of the quantity $\frac{f_i}{N}$ as $N \to \infty$. Essentially, we define the probability p_i as

$$p_i = \lim_{N\to\infty} \frac{f_i}{N}$$

Replacing $\frac{f_i}{N}$ with the probability p_i leads to the following definition of the mean or **expectation** of the discrete random variable X.

Key Point 2

The Expectation of a Discrete Random Variable

Let X be a random variable with values $x_1, x_2, \ldots, x_n$. Let the probability that X takes the value x_i (i.e. $\mathrm{P}(X = x_i)$) be denoted by p_i. The mean or **expected value** or **expectation** of X, which is written $\mathrm{E}(X)$ is defined as:

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i \, \mathrm{P}(X = x_i) = p_1x_1 + p_2x_2 + \cdots + p_nx_n$$

The symbol μ is sometimes used to denote $\mathrm{E}(X)$.

The expectation $\mathrm{E}(X)$ of X is the value of X which we expect on average. In a similar way we can write down the expected value of the function $g(X)$ as $\mathrm{E}[g(X)]$, the value of $g(X)$ we expect on average. We have

$$\mathsf{E}[g(X)] = \sum_{i}^{n} g(x_i) f(x_i)$$

In particular if $g(X) = X^2$, we obtain $\qquad \mathsf{E}[X^2] = \sum_{i}^{n} x_i^2 f(x_i)$

The variance is usually written as σ^2. For a frequency distribution it is:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \mu)^2 \qquad \text{where } \mu \text{ is the mean value}$$

and can be expanded and 'simplified' to appear as:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} f_i x_i^2 - \mu^2$$

This is often quoted in words:

The variance is equal to the mean of the squares minus the square of the mean.

We now extend the concept of variance to a random variable.

Key Point 3

The Variance of a Discrete Random Variable

Let X be a random variable with values $x_1, x_2, \ldots, x_n$. The variance of X, which is written $\mathsf{V}(X)$ is defined by

$$\mathsf{V}(X) = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

where $\mu \equiv \mathsf{E}(X)$. We note that $\mathsf{V}(X)$ can be written in the alternative form

$$\mathsf{V}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2$$

The standard deviation σ of a random variable is $\sqrt{\mathsf{V}(X)}$.

Example 6

A traffic engineer is interested in the number of vehicles reaching a particular crossroads during periods of relatively low traffic flow. The engineer finds that the number of vehicles X reaching the crossroads per minute is governed by the probability distribution:

x	0	1	2	3	4
$\mathsf{P}(X=x)$	0.37	0.39	0.19	0.04	0.01

(a) Calculate the expected value, the variance and the standard deviation of the random variable X.

(b) Graph the probability distribution $\mathsf{P}(X=x)$ and the corresponding cumulative probability distribution $F(x) = \sum_{x_i \leq x} \mathsf{P}(X = x_i)$.

Solution

(a) The expectation, variance and standard deviation and cumulative probability values are calculated as follows:

x	x^2	$\mathsf{P}(X=x)$	$F(x)$
0	0	0.37	0.37
1	1	0.39	0.76
2	4	0.19	0.95
3	9	0.04	0.99
4	16	0.01	1.00

$$\begin{aligned}\mathsf{E}(X) &= \sum_{x=0}^{4} x\mathsf{P}(X=x)\\ &= 0 \times 0.37 + 1 \times 0.39 + 2 \times 0.19 + 3 \times 0.04 + 4 \times 0.01\\ &= 0.93\end{aligned}$$

$$\begin{aligned}\mathsf{V}(X) &= \mathsf{E}(X^2) - [\mathsf{E}(X)]^2\\ &= \sum_{x=0}^{4} x^2\mathsf{P}(X=x) - \left[\sum_{x=0}^{4} x\mathsf{P}(X=x)\right]^2\\ &= 0 \times 0.37 + 1 \times 0.39 + 4 \times 0.19 + 9 \times 0.04 + 16 \times 0.01 - (0.93)^2\\ &= 0.8051\end{aligned}$$

The standard deviation is given by $\sigma = \sqrt{\mathsf{V}(X)} = 0.8973$

Solution (contd.)

(b)

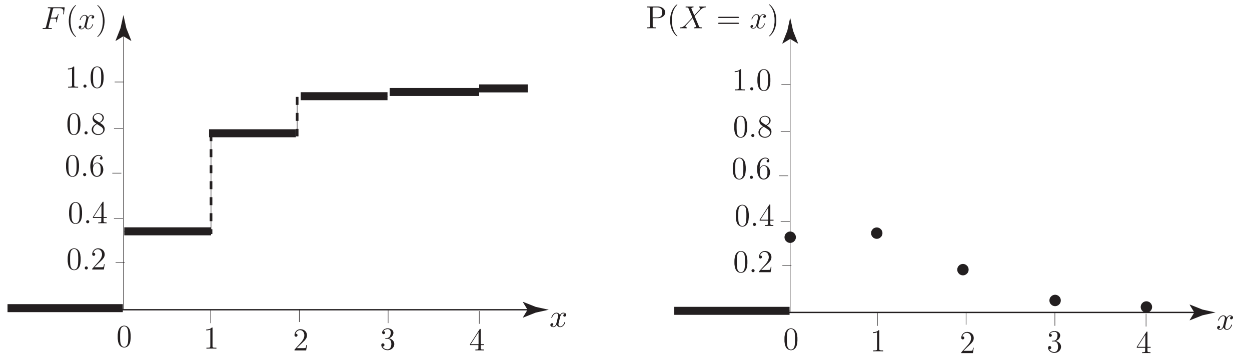

Figure 2

Task

Find the expectation, variance and standard deviation of the number of Heads in the three-coin toss experiment.

Your solution

Answer

$$\begin{aligned}
\mathsf{E}(X) &= \frac{1}{8}\times 0 + \frac{3}{8}\times 1 + \frac{3}{8}\times 2 + \frac{1}{8}\times 3 = \frac{12}{8} \\
\sum p_i x_i^2 &= \frac{1}{8}\times 0^2 + \frac{3}{8}\times 1^2 + \frac{3}{8}\times 2^2 + \frac{1}{8}\times 3^2 \\
&= \frac{1}{8}\times 0 + \frac{3}{8}\times 1 + \frac{3}{8}\times 4 + \frac{1}{8}\times 9 = 3 \\
\mathsf{V}(X) &= 3 - 2.25 = 0.75 = \frac{3}{4} \\
\sigma &= \frac{\sqrt{3}}{2}
\end{aligned}$$

Exercises

1. A machine is operated by two workers. There are sixteen workers available. How many possible teams of two workers are there?

2. A factory has 52 machines. Two of these have been given an experimental modification. In the first week after this modification, problems are reported with thirteen of the machines. What is the probability that both of the modified machines are among the thirteen with problems assuming that all machines are equally likely to give problems,?

3. A factory has 52 machines. Four of these have been given an experimental modification. In the first week after this modification, problems are reported with thirteen of the machines. What is the probability that exactly two of the modified machines are among the thirteen with problems assuming that all machines are equally likely to give problems?

4. A random number generator produces sequences of independent digits, each of which is as likely to be any digit from 0 to 9 as any other. If X denotes any single digit, find $\mathsf{E}(X)$.

5. A hand-held calculator has a clock cycle time of 100 nanoseconds; these are positions numbered $0, 1, \ldots, 99$. Assume a flag is set during a particular cycle at a random position. Thus, if X is the position number at which the flag is set.

   $$\mathsf{P}(X = k) = \frac{1}{100} \qquad k = 0, 1, 2, \ldots, 99.$$

   Evaluate the average position number $\mathsf{E}(X)$, and σ, the standard deviation.

   (Hint: The sum of the first k integers is $k(k+1)/2$ and the sum of their squares is: $k(k+1)(2k+1)/6$.)

6. Concentric circles of radii 1 cm and 3 cm are drawn on a circular target radius 5 cm. A darts player receives 10, 5 or 3 points for hitting the target inside the smaller circle, middle annular region and outer annular region respectively. The player has only a 50-50 chance of hitting the target at all but if he does hit it he is just as likely to hit any one point on it as any other. If $X =$ 'number of points scored on a single throw of a dart' calculate the expected value of X.

Answers

1. The required number is

$$\binom{16}{2} = \frac{16 \times 15}{2 \times 1} = 120.$$

2. There are

$$\binom{52}{13}$$

possible different selections of 13 machines and all are equally likely. There is only

$$\binom{2}{2} = 1$$

way to pick two machines from those which were modified but there are

$$\binom{50}{11}$$

different choices for the 11 other machines with problems so this is the number of possible selections containing the 2 modified machines.

Hence the required probability is

$$\begin{aligned}\frac{\binom{2}{2}\binom{50}{11}}{\binom{52}{13}} &= \frac{\binom{50}{11}}{\binom{52}{13}} \\ &= \frac{50!/(11!39!)}{52!/(13!39!)} \\ &= \frac{50!13!}{52!11!} \\ &= \frac{13 \times 12}{52 \times 51} \approx 0.0588\end{aligned}$$

Alternatively, let S be the event "first modified machine is in the group of 13" and C be the event "second modified machine is in the group of 13". Then the required probability is

$$\mathsf{P}(S) \times \mathsf{P}(C \mid S) = \frac{13}{52} \times \frac{12}{51}.$$

Answers

3. There are $\begin{pmatrix} 52 \\ 13 \end{pmatrix}$ different selections of 13, $\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ different choices of two modified machines and $\begin{pmatrix} 48 \\ 11 \end{pmatrix}$ different choices of 11 non-modified machines.

Thus the required probability is

$$\frac{\begin{pmatrix} 4 \\ 2 \end{pmatrix}\begin{pmatrix} 48 \\ 11 \end{pmatrix}}{\begin{pmatrix} 52 \\ 13 \end{pmatrix}} = \frac{(4!/2!2!)(48!/11!37!)}{(52!/13!39!)}$$

$$= \frac{4!48!13!39!}{52!2!2!11!37!}$$

$$= \frac{4 \times 3 \times 13 \times 12 \times 39 \times 38}{52 \times 51 \times 50 \times 49 \times 2} \approx 0.2135$$

Alternatively, let $I(i)$ be the event "modified machine i is in the group of 13" and $O(i)$ be the negation of this, for $i = 1, 2, 3, 4$. The number of choices of two modified machines is

$$\begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

so the required probability is

$$\begin{pmatrix} 4 \\ 2 \end{pmatrix} \mathsf{P}\{I(1)\} \times \mathsf{P}\{I(2) \mid I(1)\} \times \mathsf{P}\{O(3) \mid I(1), I(2)\} \times \mathsf{P}\{O(4) \mid I(1)I(2)O(3)\}$$

$$= \begin{pmatrix} 4 \\ 2 \end{pmatrix} \frac{13}{52} \times \frac{12}{51} \times \frac{39}{50} \times \frac{38}{49}$$

$$= \frac{4 \times 3 \times 13 \times 12 \times 39 \times 38}{52 \times 51 \times 50 \times 49 \times 2}$$

4.

x	0	1	2	3	4	5	6	7	8	9
$\mathsf{P}(X = x)$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$	$^1/_{10}$

$$\mathsf{E}(X) = \frac{1}{10}\{0 + 1 + 2 + 3 + \ldots + 9\} = 4.5$$

Answers

5. Same as Q.4 but with 100 positions

$$\mathsf{E}(X) = \frac{1}{100}\{0+1+2+3+\ldots+99\} = \frac{1}{100}\left[\frac{99(99+1))}{2}\right] = 49.5$$

$\sigma^2 =$ mean of squares $-$ square of means

$$\therefore \quad \sigma^2 = \frac{1}{100}[1^2+2^2+\ldots+99^2] - (49.5)^2$$
$$= \frac{1}{100}\frac{[99(100)(199)]}{6} - 49.5^2 = 833.25$$

so the standard deviation is $\sigma = \sqrt{833.25} = 28.87$

6. X can take 4 values 0, 3, 5 or 10

$\mathsf{P}(X=0) = 0.5$ [only 50/50 chance of hitting target]

The probability that a particular points score is obtained is related to the areas of the annular regions which are, from the centre: $\pi,\ (9\pi-\pi) = 8\pi,\ (25\pi - 9\pi) = 16\pi$

$$\mathsf{P}(X=3\ ) = \mathsf{P}[(3 \text{ is scored}) \cap (\text{target is hit})]$$
$$= \mathsf{P}(3 \text{ is scored} \,|\, \text{target is hit}) \times \mathsf{P}(\text{target is hit})$$
$$= \frac{16\pi}{25\pi}.\frac{1}{2} = \frac{16}{50}$$
$$\mathsf{P}(X=5\ ) = \mathsf{P}(5 \text{ is scored} \,|\, \text{target is hit}) \times \mathsf{P}(\text{target is hit})$$
$$= \frac{8\pi}{25\pi}.\frac{1}{2} = \frac{8}{50}$$
$$\mathsf{P}(X=10) = \mathsf{P}(10 \text{ is scored} \,|\, \text{target is hit}) \times \mathsf{P}(\text{target is hit})$$
$$= \frac{\pi}{25\pi}.\frac{1}{2} = \frac{1}{50}$$

x	0	3	5	10
$\mathsf{P}(X=x)$	$^{25}/_{50}$	$^{16}/_{50}$	$^{8}/_{50}$	$^{1}/_{50}$

$$\therefore \quad \mathsf{E}(X) = \frac{48+40+10}{50} = 1.96.$$

The Binomial Distribution

37.2

Introduction

A situation in which an experiment (or trial) is repeated a fixed number of times can be modelled, under certain assumptions, by the binomial distribution. Within each trial we focus attention on a particular outcome. If the outcome occurs we label this as a success. The binomial distribution allows us to calculate the probability of observing a certain number of successes in a given number of trials.

You should note that the term 'success' (and by implication 'failure') are simply labels and as such might be misleading. For example counting the number of defective items produced by a machine might be thought of as counting successes if you are looking for defective items! Trials with two possible outcomes are often used as the building blocks of random experiments and can be useful to engineers. Two examples are:

1. A particular mobile phone link is known to transmit 6% of 'bits' of information in error. As an engineer you might need to know the probability that two bits out of the next ten transmitted are in error.

2. A machine is known to produce, on average, 2% defective components. As an engineer you might need to know the probability that 3 items are defective in the next 20 produced.

The binomial distribution will help you to answer such questions.

Prerequisites

Before starting this Section you should ...

- understand the concepts of probability

Learning Outcomes

On completion you should be able to ...

- recognise and use the formula for binomial probabilities
- state the assumptions on which the binomial model is based

1. The binomial model

We have introduced random variables from a general perspective and have seen that there are two basic types: discrete and continuous. We examine four particular examples of distributions for random variables which occur often in practice and have been given special names. They are the **binomial** distribution, the **Poisson** distribution, the **Hypergeometric** distribution and the **Normal** distribution. The first three are distributions for discrete random variables and the fourth is for a continuous random variable. In this Section we focus attention on the binomial distribution.

The binomial distribution can be used in situations in which a given experiment (often referred to, in this context, as a **trial**) is repeated a number of times. For the binomial model to be applied the following four criteria must be satisfied:

- the trial is carried out a fixed number of times n
- the outcomes of each trial can be classified into two 'types' conventionally named 'success' or 'failure'
- the probability p of success remains constant for each trial
- the individual trials are independent of each other.

For example, if we consider throwing a coin 7 times what is the probability that exactly 4 Heads occur? This problem can be modelled by the binomial distribution since the four basic criteria are assumed satisfied as we see.

- here the trial is 'throwing a coin' which is carried out 7 times
- the occurrence of Heads on any given trial (i.e. throw) may be called a 'success' and Tails called a 'failure'
- the probability of success is $p = \frac{1}{2}$ and remains constant for each trial
- each throw of the coin is independent of the others.

The reader will be able to complete the solution to this example once we have constructed the general binomial model.

The following two scenarios are typical of those met by engineers. The reader should check that the criteria stated above are met by each scenario.

1. An electronic product has a total of 30 integrated circuits built into it. The product is capable of operating successfully only if at least 27 of the circuits operate properly. What is the probability that the product operates successfully if the probability of any integrated circuit failing to operate is 0.01?
2. Digital communication is achieved by transmitting information in "bits". Errors do occur in data transmissions. Suppose that the number of bits in error is represented by the random variable X and that the probability of a communication error in a bit is 0.001. If at most 2 errors are present in a 1000 bit transmission, the transmission can be successfully decoded. If a 1000 bit message is transmitted, find the probability that it can be successfully decoded.

Before developing the *general* binomial distribution we consider the following examples which, as you will soon recognise, have the basic characteristics of a binomial distribution.

Example 7

In a box of floppy discs it is known that 95% will work. A sample of three of the discs is selected at random.
Find the probability that (a) none (b) 1, (c) 2, (d) all 3 of the sample will work.

Solution

Let the event {the disc works} be W and the event {the disc fails} be F. The probability that a disc will work is denoted by $\mathsf{P}(W)$ and the probability that a disc will fail is denoted by $\mathsf{P}(F)$. Then $\mathsf{P}(W) = 0.95$ and $\mathsf{P}(F) = 1 - \mathsf{P}(W) = 1 - 0.95 = 0.05$.

(a) The probability that none of the discs works equals the probability that all 3 discs fail. This is given by:

$$\begin{aligned} \mathsf{P}(\text{none work}) = \mathsf{P}(FFF) &= \mathsf{P}(F)\times\mathsf{P}(F)\times\mathsf{P}(F) \qquad \text{as the events are independent} \\ &= 0.05\times0.05\times0.05 = 0.05^3 = 0.000125 \end{aligned}$$

(b) If only one disc works then you could select the three discs in the following orders

$$(FFW) \text{ or } (FWF) \text{ or } (WFF) \qquad \text{hence}$$

$$\begin{aligned} \mathsf{P}(\text{one works}) &= \mathsf{P}(FFW)+\mathsf{P}(FWF)+\mathsf{P}(WFF) \\ &= \mathsf{P}(F)\times\mathsf{P}(F)\times\mathsf{P}(W)+\mathsf{P}(F)\times\mathsf{P}(W)\times\mathsf{P}(F)+\mathsf{P}(W)\times\mathsf{P}(F)\times\mathsf{P}(F) \\ &= (0.05\times0.05\times0.95)+(0.05\times0.95\times0.05)+(0.95\times0.05\times0.05) \\ &= 3\times(0.05)^2\times0.95 = 0.007125 \end{aligned}$$

(c) If 2 discs work you could select them in order

$$(FWW) \text{ or } (WFW) \text{ or } (WWF) \qquad \text{hence}$$

$$\begin{aligned} \mathsf{P}(\text{two work}) &= \mathsf{P}(FWW)+\mathsf{P}(WFW)+\mathsf{P}(WWF) \\ &= \mathsf{P}(F)\times\mathsf{P}(W)\times\mathsf{P}(W)+\mathsf{P}(W)\times\mathsf{P}(F)\times\mathsf{P}(W)+\mathsf{P}(W)\times\mathsf{P}(W)\times\mathsf{P}(F) \\ &= (0.05\times0.95\times0.95)+(0.95\times0.05\times0.95)+(0.95\times0.95\times0.05) \\ &= 3\times(0.05)\times(0.95)^2 = 0.135375 \end{aligned}$$

(d) The probability that all 3 discs work is given by $\mathsf{P}(WWW) = 0.95^3 = 0.857375$.

Notice that since the 4 outcomes we have dealt with are *all possible outcomes* of selecting 3 discs, the probabilities should add up to 1. It is an easy check to verify that they do.

One of the most important assumptions above is that of **independence**.The probability of selecting a working disc remains unchanged no matter whether the previous selected disc worked or not.

Example 8

A worn machine is known to produce 10% defective components. If the random variable X is the number of defective components produced in a run of 3 components, find the probabilities that X takes the values 0 to 3.

Solution

Assuming that the production of components is independent and that the probability $p = 0.1$ of producing a defective component remains constant, the following table summarizes the production run. We let G represent a *good* component and let D represent a *defective* component.
Note that since we are only dealing with two possible outcomes, we can say that the probability q of the machine producing a good component is $1 - 0.1 = 0.9$. More generally, we know that $q+p = 1$ if we are dealing with a binomial distribution.

Outcome	Value of X	Probability of Occurrence
GGG	0	$(0.9)(0.9)(0.9) = (0.9)^3$
GGD	1	$(0.9)(0.9)(0.1) = (0.9)^2(0.1)$
GDG	1	$(0.9)(0.1)(0.9) = (0.9)^2(0.1)$
DGG	1	$(0.1)(0.9)(0.9) = (0.9)^2(0.1)$
DDG	2	$(0.1)(0.1)(0.9) = (0.9)(0.1)^2$
DGD	2	$(0.1)(0.9)(0.1) = (0.9)(0.1)^2$
GDD	2	$(0.9)(0.1)(0.1) = (0.9)(0.1)^2$
DDD	3	$(0.1)(0.1)(0.1) = (0.1)^3$

From this table it is easy to see that

$$\mathsf{P}(X = 0) = (0.9)^3$$

$$\mathsf{P}(X = 1) = 3 \times (0.9)^2(0.1)$$

$$\mathsf{P}(X = 2) = 3 \times (0.9)(0.1)^2$$

$$\mathsf{P}(X = 3) = (0.1)^3$$

Clearly, a pattern is developing. In fact you may have already realized that the probabilities we have found are just the terms of the expansion of the expression $(0.9 + 0.1)^3$ since

$$(0.9 + 0.1)^3 = (0.9)^3 + 3 \times (0.9)^2(0.1) + 3 \times (0.9)(0.1)^2 + (0.1)^3$$

We now develop the binomial distribution from a more general perspective. If you find the theory getting a bit heavy simply refer back to this example to help clarify the situation.
First we shall find it convenient to denote the probability of failure on a trial, which is $1 - p$, by q, that is:

$$q = 1 - p.$$

What we shall do is to calculate probabilities of the number of 'successes' occurring in n trials, beginning with $n = 1$.

$n = 1$ With only one trial we can observe either 1 success (with probability p) or 0 successes (with probability q).

$\underline{\boldsymbol{n=2}}$ Here there are 3 possibilities: We can observe 2, 1 or 0 successes. Let S denote a success and F denote a failure. So a failure followed by a success would be denoted by FS whilst two failures followed by one success would be denoted by FFS and so on.
Then

$$\mathsf{P}(\text{2 successes in 2 trials}) = \mathsf{P}(SS) = \mathsf{P}(S)\mathsf{P}(S) = p^2$$

(where we have used the assumption of independence between trials and hence multiplied probabilities). Now, using the usual rules of basic probability, we have:

$$\mathsf{P}(\text{1 success in 2 trials}) = \mathsf{P}[(SF) \cup (FS)] = \mathsf{P}(SF) + \mathsf{P}(FS) = pq + qp = 2pq$$

$$\mathsf{P}(\text{0 successes in 2 trials}) = \mathsf{P}(FF) = \mathsf{P}(F)\mathsf{P}(F) = q^2$$

The three probabilities we have found $-$ $q^2,\ 2qp,\ p^2$ $-$ are in fact the terms which arise in the binomial expansion of $(q+p)^2 = q^2 + 2qp + p^2$. We also note that since $q = 1 - p$ the probabilities sum to 1 (as we should expect):

$$q^2 + 2qp + p^2 = (q+p)^2 = ((1-p)+p)^2 = 1$$

Task

List the outcomes for the binomial model for the case $n = 3$, calculate their probabilities and display the results in a table.

Your solution

Answer

{three successes, two successes, one success, no successes}

Three successes occur only as SSS with probability p^3.

Two successes can occur as SSF with probability (p^2q), as SFS with probability (pqp) or as FSS with probability (qp^2).

These are mutually exclusive events so the combined probability is the sum $3p^2q$.

Similarly, we can calculate the other probabilities and obtain the following table of results.

Number of successes	3	2	1	0
Probability	p^3	$3p^2q$	$3pq^2$	q^3

Note that the probabilities you have obtained:

$$q^3,\ 3q^2p,\ 3qp^2,\ p^3$$

are the terms which arise in the binomial expansion of $(q+p)^3 = q^3 + 3q^2p + 3qp^2 + p^3$

Repeat the previous Task for the binomial model for the case with $n = 4$.

Your solution

Answer

Number of successes	4	3	2	1	0
Probability	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4

Again we explore the connection between the probabilities and the terms in the binomial expansion of $(q+p)^4$. Consider this expansion

$$(q+p)^4 = q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4$$

Then, for example, the term $4p^3q$, is the probability of 3 successes in the four trials. These successes can occur anywhere in the four trials and there must be one failure hence the p^3 and q components which are multiplied together. The remaining part of this term, 4, is the number of ways of selecting three objects from 4.

Similarly there are ${}^4C_2 = \dfrac{4!}{2!2!} = 6$ ways of selecting two objects from 4 so that the coefficient 6 combines with p^2 and q^2 to give the probability of two successes (and hence two failures) in four trials.

The approach described here can be extended for any number n of trials.

Key Point 4

The Binomial Probabilities

Let X be a discrete random variable, being the number of successes occurring in n independent trials of an experiment. If X is to be described by the binomial model, the probability of exactly r successes in n trials is given by

$$\text{P}(X = r) = {}^nC_r p^r q^{n-r}.$$

Here there are r successes (each with probability p), $n - r$ failures (each with probability q) and ${}^nC_r = \dfrac{n!}{r!(n-r)!}$ is the number of ways of placing the r successes among the n trials.

Notation

If a random variable X follows a binomial distribution in which an experiment is repeated n times each with probability p of success then we write $X \sim B(n, p)$.

Example 9

A worn machine is known to produce 10% defective components. If the random variable X is the number of defective components produced in a run of 4 components, find the probabilities that X takes the values 0 to 4.

Solution

From Example 8, we know that the probabilities required are the terms of the expansion of the expression:

$(0.9 + 0.1)^4$ so $X \sim B(4, 0.1)$

Hence the required probabilities are (using the general formula with $n = 4$ and $p = 0.1$)

$$\text{P}(X = 0) = (0.9)^4 = 0.6561$$

$$\text{P}(X = 1) = 4(0.9)^3(0.1) = 0.2916$$

$$\text{P}(X = 2) = \frac{4 \times 3}{1 \times 2}(0.9)^2(0.1)^2 = 0.0486$$

$$\text{P}(X = 3) = \frac{4 \times 3 \times 2}{1 \times 2 \times 3}(0.9)(0.1)^3 = 0.0036$$

$$\text{P}(X = 4) = (0.1)^4 = 0.0001$$

Also, since we are using the expansion of $(0.9 + 0.1)^4$, the probabilities should sum to 1, This is a useful check on your arithmetic when you are using a binomial distribution.

Example 10

In a box of switches it is known 10% of the switches are faulty. A technician is wiring 30 circuits, each of which needs one switch. What is the probability that (a) all 30 work, (b) at most 2 of the circuits do not work?

Solution

The answers involve binomial distributions because there are only two states for each circuit - it either works or it doesn't work.

A trial is the operation of testing each circuit.
A success is that it works. We are given P(success) $= p = 0.9$
Also we have the number of trials $n = 30$

Applying the binomial distribution $\mathsf{P}(X = r) = {}^nC_r p^r (1-p)^{n-r}$.

(a) Probability that all 30 work is $\mathsf{P}(X = 30) = {}^{30}C_{30}(0.9)^{30}(0.1)^0 = 0.04239$

(b) The statement that "at most 2 circuits do not work" implies that 28, 29 or 30 work. That is $X \geq 28$

$$
\begin{aligned}
\mathsf{P}(X \geq 28) &= \mathsf{P}(X = 28) + \mathsf{P}(X = 29) + \mathsf{P}(X = 30) \\
\mathsf{P}(X = 30) &= {}^{30}C_{30}(0.9)^{30}(0.1)^0 = 0.04239 \\
\mathsf{P}(X = 29) &= {}^{30}C_{29}(0.9)^{29}(0.1)^1 = 0.14130 \\
\mathsf{P}(X = 28) &= {}^{30}C_{28}(0.9)^{28}(0.1)^2 = 0.22766
\end{aligned}
$$

Hence $\mathsf{P}(X \geq 28) = 0.41135$

Example 11

A University Engineering Department has introduced a new software package called SOLVIT. To save money, the University's Purchasing Department has negotiated a bargain price for a 4-user licence that allows only four students to use SOLVIT at any one time. It is estimated that this should allow 90% of students to use the package when they need it. The Students' Union has asked for more licences to be bought since engineering students report having to queue excessively to use SOLVIT. As a result the Computer Centre monitors the use of the software. Their findings show that on average 20 students are logged on at peak times and 4 of these want to use SOLVIT. Was the Purchasing Department's estimate correct?

Solution

P(student wanted to use SOLVIT) $= \dfrac{4}{20} = 0.2$

Let X be the number of students wanting to use SOLVIT at any one time, then

$$\begin{aligned}
\mathsf{P}(X=0) &= {}^{20}C_0(0.2)^0(0.8)^{20} = 0.0115\\
\mathsf{P}(X=1) &= {}^{20}C_1(0.2)^1(0.8)^{19} = 0.0576\\
\mathsf{P}(X=2) &= {}^{20}C_2(0.2)^2(0.8)^{18} = 0.1369\\
\mathsf{P}(X=3) &= {}^{20}C_3(0.2)^3(0.8)^{17} = 0.2054\\
\mathsf{P}(X=4) &= {}^{20}C_4(0.2)^4(0.8)^{16} = 0.2182
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathsf{P}(X\leq 4) &= \mathsf{P}(X=0)+\mathsf{P}(X=1)+\mathsf{P}(X=2)+\mathsf{P}(X=3)+\mathsf{P}(X=4)\\
&= 0.01152+0.0576+0.1369+0.2054+0.2182\\
&= 0.61862
\end{aligned}$$

The probability that more than 4 students will want to use SOLVIT is

$$\mathsf{P}(X>4) = 1-\mathsf{P}(X\leq 4) = 0.38138$$

That is, 38% of the time there will be more than 4 students wanting to use the software. The Purchasing Department has grossly overestimated the availability of the software on the basis of a 4-user licence.

Task

Using the binomial model, and assuming that a success occurs with probability $\frac{1}{5}$ in each trial, find the probability that in 6 trials there are
(a) 0 successes (b) 3 successes (c) 2 failures.

Let X be the number of successes in 6 independent trials.

Your solution

(a) $\mathsf{P}(X = 0) =$

Answer

In each case $p = \dfrac{1}{5}$ and $q = 1 - p = \dfrac{4}{5}$.
Here $r = 0$ and

$$\mathsf{P}(X = 0) = q^6 = \left(\frac{4}{5}\right)^6 = \frac{4096}{15625} \approx 0.262$$

Your solution

(b) $\mathsf{P}(X = 3) =$

Answer

$r = 3$ and $\quad \mathsf{P}(X = 3) = {}^6C_3 p^3 q^3 = \dfrac{6 \times 5 \times 4}{1 \times 2 \times 3} \times \left(\dfrac{1}{5}\right)^3 \times \left(\dfrac{4}{5}\right)^3 = \dfrac{20 \times 64}{5^6} = \dfrac{12 \times 80}{15625} = 0.0819$

Your solution

(c) $\mathsf{P}(X = 4) =$

Answer

Here $r = 4$ and $\mathsf{P}(X = 4) = {}^6C_4 p^4 q^2 = \dfrac{6 \times 5}{1 \times 2} \times \left(\dfrac{1}{5}\right)^4 \times \left(\dfrac{4}{5}\right)^2 = \dfrac{15 \times 4^2}{5^6} = \dfrac{240}{15625} = 0.01536$

2. Expectation and variance of the binomial distribution

For a binomial distribution $X \sim B(n, p)$, the mean and variance, as we shall see, have a simple form. While we will not prove the formulae in general terms - the algebra can be rather tedious - we will illustrate the results for cases involving small values of n.

The case $n = 2$

Essentially, we have a random variable X which follows a binomial distribution $X \sim B(2, p)$ so that the values taken by X (and X^2 - needed to calculate the variance) are shown in the following table:

x	x^2	$\mathsf{P}(X = x)$	$x\mathsf{P}(X = x)$	$x^2\mathsf{P}(X = x)$
0	0	q^2	0	0
1	1	$2qp$	$2qp$	$2qp$
2	4	p^2	$2p^2$	$4p^2$

We can now calculate the mean of this distribution:

$$\mathsf{E}(X) = \sum x\mathsf{P}(X = x) = 0 + 2qp + 2p^2 = 2p(q + p) = 2p \qquad \text{since } q + p = 1$$

Similarly, the variance $V(X)$ is given by

$$V(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = 0 + 2qp + 4p^2 - (2p)^2 = 2qp$$

Calculate the mean and variance of a random variable X which follows a binomial distribution $X \sim B(3, p)$.

Your solution

Answer

The table of values appropriate to this case is:

x	x^2	$\mathsf{P}(X=x)$	$x\mathsf{P}(X=x)$	$x^2\mathsf{P}(X=x)$
0	0	q^3	0	0
1	1	$3q^2p$	$3q^2p$	$3q^2p$
2	4	$3qp^2$	$6qp^2$	$12qp^2$
3	9	p^3	$3p^3$	$9p^3$

Hence $\mathsf{E}(X) = \sum x\mathsf{P}(X=x) = 0 + 3q^2p + 6qp^2 + 3p^3 = 3p(q+p)^2 = 3p$ since $q+p=1$

$$\begin{aligned} V(X) &= \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 \\ &= 0 + 3q^2p + 12qp^2 + 9p^3 - (3p)^2 \\ &= 3p(q^2 + 4qp + 3p^2 - 3p) \\ &= 3p((1-p)^2 + 4(1-p)p + 3p^2 - 3p) \\ &= 3p(1 - 2p + p^2 + 4p - 4p^2 + 3p^2 - 3p) = 3p(1-p) = 3pq \end{aligned}$$

From the results given above, it is reasonable to asert the following result in Key Point 5.

Key Point 5

Expectation and Variance of the Binomial Distribution

If a random variable X which can assume the values $0, 1, 2, 3, \ldots, n$ follows a binomial distribution $X \sim B(n,p)$ so that

$$\mathsf{P}(X=r) = {}^nC_r p^r q^{n-r} = {}^nC_r p^r (1-p)^{n-r}$$

then the expectation and variance of the distribution are given by the formulae

$$\mathsf{E}(X) = np \qquad \text{and} \qquad V(X) = np(1-p) = npq$$

Task

A die is thrown repeatedly 36 times in all. Find $\mathsf{E}(X)$ and $V(X)$ where X is the number of sixes obtained.

Your solution

Answer

Consider the occurrence of a six, with X being the number of sixes thrown in 36 trials.

The random variable X follows a binomial distribution. (Why? Refer to page 18 for the criteria if necessary). A trial is the operation of throwing a die. A success is the occurrence of a 6 on a particular trial, so $p = \frac{1}{6}$. We have $n = 36,\ p = \frac{1}{6}$ so that

$$\mathsf{E}(X) = np = 36 \times \frac{1}{6} = 6 \qquad \text{and} \qquad V(X) = npq = 36 \times \frac{1}{6} \times \frac{5}{6} = 5.$$

Hence the standard deviation is $\sigma = \sqrt{5} \simeq 2.236$.

$\mathsf{E}(X) = 6$ implies that in 36 throws of a fair die we would expect, on average, to see 6 sixes. This makes perfect sense, of course.

Exercises

1. The probability that a mountain-bike rider travelling along a certain track will have a tyre burst is 0.05. Find the probability that among 17 riders:

   (a) exactly one has a burst tyre

   (b) at most three have a burst tyre

   (c) two or more have burst tyres.

2. (a) A transmission channel transmits zeros and ones in strings of length 8, called 'words'. Possible distortion may change a one to a zero or vice versa; assume this distortion occurs with probability .01 for each digit, independently. An error-correcting code is employed in the construction of the word such that the receiver can deduce the word correctly if at most one digit is in error. What is the probability the word is decoded incorrectly?

   (b) Assume that a word is a sequence of 10 zeros or ones and, as before, the probability of incorrect transmission of a digit is .01. If the error-correcting code allows correct decoding of the word if no more than two digits are incorrect, compute the probability that the word is decoded correctly.

3. An examination consists of 10 multi-choice questions, in each of which a candidate has to deduce which one of five suggested answers is correct. A completely unprepared student guesses each answer completely randomly. What is the probability that this student gets 8 or more questions correct? Draw the appropriate moral!

4. The probability that a machine will produce all bolts in a production run within specification is 0.998. A sample of 8 machines is taken at random. Calculate the probability that

   (a) all 8 machines, (b) 7 or 8 machines, (c) at least 6 machines

   will produce all bolts within specification

5. The probability that a machine develops a fault within the first 3 years of use is 0.003. If 40 machines are selected at random, calculate the probability that 38 or more will not develop any faults within the first 3 years of use.

6. A computer installation has 10 terminals. Independently, the probability that any one terminal will require attention during a week is 0.1. Find the probabilities that

   (a) 0, (b), 1 (c) 2, (d) 3 or more, terminals will require attention during the next week.

7. The quality of electronic chips is checked by examining samples of 5. The frequency distribution of the number of defective chips per sample obtained when 100 samples have been examined is:

No. of defectives	0	1	2	3	4	5
No. of samples	47	34	16	3	0	0

   Calculate the proportion of defective chips in the 500 tested. Assuming that a binomial distribution holds, use this value to calculate the expected frequencies corresponding to the observed frequencies in the table.

Exercises continued

8. In a large school, 80% of the pupils like mathematics. A visitor to the school asks each of 4 pupils, chosen at random, whether they like mathematics.

(a) Calculate the probabilities of obtaining an answer *yes* from 0, 1, 2, 3, 4 of the pupils

(b) Find the probability that the visitor obtains the answer *yes* from at least 2 pupils:

(i) when the number of pupils questioned remains at 4

(ii) when the number of pupils questioned is increased to 8.

9. A machine has two drive belts, one on the left and one on the right. From time to time the drive belts break. When one breaks the machine is stopped and both belts are replaced. Details of n consecutive breakages are recorded. Assume that the left and right belts are equally likely to break first. Let X be the number of times the break is on the left.

(a) How many possible different sequences of "left" and "right" are there?

(b) How many of these sequences contain exactly j "lefts"?

(c) Find an expression, in terms of n and j, for the probability that $X = j$.

(d) Let $n = 6$. Find the probability distribution of X.

10. A machine is built to make mass-produced items. Each item made by the machine has a probability p of being defective. Given the value of p, the items are independent of each other. Because of the way in which the machines are made, p could take one of several values. In fact $p = X/100$ where X has a discrete uniform distribution on the interval $[0, 5]$. The machine is tested by counting the number of items made before a defective is produced. Find the conditional probability distribution of X given that the first defective item is the thirteenth to be made.

11. Seven batches of articles are manufactured. Each batch contains ten articles. Each article has, independently, a probability of 0.1 of being defective. Find the probability that there is at least one defective article

(a) in exactly four of the batches,

(b) in four or more of the batches.

12. A service engineer is can be called out for maintenance on the photocopiers in the offices of four large companies, A, B, C and D. On any given week there is a probability of 0.1 that he will be called to each of these companies. The event of being called to one company is independent of whether or not he is called to any of the others.

(a) Find the probability that, on a particular day,

(i) he is called to all four companies,

(ii) he is called to at least three companies,

(iii) he is called to all four given that he is called to at least one,

(iv) he is called to all four given that he is called to Company A.

(b) Find the expected value and variance of the number of these companies which call the engineer on a given day.

Exercises continued

13. There are five machines in a factory. Of these machines, three are working properly and two are defective. Machines which are working properly produce articles each of which has independently a probability of 0.1 of being imperfect. For the defective machines this probability is 0.2. A machine is chosen at random and five articles produced by the machine are examined. What is the probability that the machine chosen is defective given that, of the five articles examined, two are imperfect and three are perfect?

14. A company buys mass-produced articles from a supplier. Each article has a probability p of being defective, independently of other articles. If the articles are manufactured correctly then $p = 0.05$. However, a cheaper method of manufacture can be used and this results in $p = 0.1$.

(a) Find the probability of observing exactly three defectives in a sample of twenty articles

(i) given that $p = 0.05$

(ii) given that $p = 0.1$.

(b) The articles are made in large batches. Unfortunately batches made by both methods are stored together and are indistinguishable until tested, although all of the articles in any one batch will be made by the same method. Suppose that a batch delivered to the company has a probability of 0.7 of being made by the correct method. Find the conditional probability that such a batch is correctly manufactured given that, in a sample of twenty articles from the batch, there are exactly three defectives.

(c) The company can either accept or reject a batch. Rejecting a batch leads to a loss for the company of £150. Accepting a batch which was manufactured by the cheap method will lead to a loss for the company of £400. Accepting a batch which was correctly manufactured leads to a profit of £500. Determine a rule for what the company should do if a sample of twenty articles contains exactly three defectives, in order to maximise the expected value of the profit (where loss is negative profit). Should such a batch be accepted or rejected?

(d) Repeat the calculation for four defectives in a sample of twenty and hence, or otherwise, determine a rule for how the company should decide whether to accept or reject a batch according to the number of defectives.

Answers

1. Binomial distribution $\mathsf{P}(X=r) = {}^nC_r p^r(1-p)^{n-r}$ where p is the probability of single 'success' which is 'tyre burst'.

   (a) $\mathsf{P}(X=1) = {}^{17}C_1(0.05)^1(0.95)^{16} = 0.3741$

   (b)

$$\begin{aligned}\mathsf{P}(X\leq 3) &= \mathsf{P}(X=0)+\mathsf{P}(X=1)+\mathsf{P}(X=2)+\mathsf{P}(X=3)\\ &= (0.95)^{17}+17(0.05)(0.95)^{16}+\frac{17\times 16}{2\times 1}(0.05)^2(0.95)^{15}\\ &\quad +\frac{17\times 16\times 15}{3\times 2\times 1}(0.05)^3(0.95)^{14} = 0.9912\end{aligned}$$

   (c) $\mathsf{P}(X\geq 2) = 1-\mathsf{P}[(X=0)\cup(X=1)] = 1-(0.95)^{17}-17(0.05)(0.95)^{16} = 0.2077$

2.

   (a) P (distortion) $= 0.01$ for each digit. This is a binomial situation in which the probability of 'success' is $0.01 = p$ and there are $n=8$ trials.
   A word is decoded incorrectly if there are two or more digits in error

$$\begin{aligned}\mathsf{P}(X\geq 2) &= 1-\mathsf{P}[(X=0)\cup(X=1)]\\ &= 1-{}^8C_0(0.99)^8-{}^8C_1(0.01)(0.99)^7 = 0.00269\end{aligned}$$

   (b) Same as (a) with $n=10$. Correct decoding if $X\leq 2$

$$\begin{aligned}\mathsf{P}(X\leq 2) &= \mathsf{P}[(X=0)\cup(X=1)\cup(X=2)]\\ &= (0.99)^{10}+10(0.01)(0.99)^9+45(0.01)^2(0.99)^8 = 0.99989\end{aligned}$$

3. Let X be a random variable 'number of answers guessed correctly' then for each question (i.e. trial) the probability of a 'success' $=\frac{1}{5}$. It is clear that X follows a binomial distribution with $n=10$ and $p=0.2$.
   P (randomly choosing correct answer) $=\frac{1}{5}$ $\quad n=10$

$$\begin{aligned}\mathsf{P}(8 \text{ or more correct}) &= \mathsf{P}[(X=8)\cup(X=9)\cup(X=10)]\\ &= {}^{10}C_8(0.2)^8(0.8)^2+{}^{10}C_9(0.2)^9(0.8)+{}^{10}C_{10}(0.2)^{10} = 0.000078\end{aligned}$$

4. (a) 0.9841 (b) 0.9999 (c) 1.0000

5. $\mathsf{P}(X\geq 38) = \mathsf{P}(X=38)+\mathsf{P}(X=39)+\mathsf{P}(X=40) = 0.00626+0.1067+0.88676 = 0.99975$

6. (a) 0.3487 (b) 0.3874 (c) 0.1937 (d) 0.0702

7. 0.15 (total defectives = 0 + 34 + 32 + 9 + 0 out of 500 tested); 44, 39, 14, 2, 0, 0

8. (a) 0.0016, 0.0256, 0.1536, 0.4096, 0.4096; (b)(i) 0.9728 (b)(ii) 0.9988

Answers

9.

(a) There are 2^n possible sequences.

(b) The number containing exactly j "lefts" is $\begin{pmatrix} n \\ j \end{pmatrix}$.

(c) $\mathsf{P}(X = j) = \begin{pmatrix} n \\ j \end{pmatrix} 2^{-n}$.

(d) With $n = 6$ the distribution of X is

j	0	1	2	3	4	5	6
$\mathsf{P}(X = j)$	0.015625	0.09375	0.234375	0.3125	0.234375	0.09375	0.015625

10. Let Y be the number of the first defective item.

$$\mathsf{P}(X = j \mid Y = 13) = \frac{\mathsf{P}(X = j) \times \mathsf{P}(Y = 13 \mid X = j)}{\sum_{i=0}^{5} \mathsf{P}(X = i) \times \mathsf{P}(Y = 13 \mid X = i)} = \frac{\mathsf{P}(Y = 13 \mid X = j)}{\sum_{i=0}^{5} \mathsf{P}(Y = 13 \mid X = i)}$$

since $\mathsf{P}(X = j) = 1/6$ for $j = 0, \ldots, 5$.

$$\mathsf{P}(Y = 13 \mid X = j) = \left(1 - \frac{X}{100}\right)^{12} \left(\frac{X}{100}\right)$$

j	$\mathsf{P}(Y = 13 \mid X = j)$	$\mathsf{P}(X = j \mid Y = 13)$
0	0.00000	0.0000
1	0.00886	0.0707
2	0.01569	0.1251
3	0.02082	0.1660
4	0.02451	0.1954
5	0.02702	0.2154
6	0.02856	0.2277
Total	0.12546	1

Answers

11.

The probability of at least one defective in a batch is $1 - 0.9^{10} = 0.6513$.
Let the probability of at least one defective in exactly j batches be p_j.

(a) $$\mathsf{p}_4 = \binom{7}{4} \left(1 - 0.9^{10}\right)^4 \left(0.9^{10}\right)^3 = 35 \times 0.6513^4 \times 0.3487^3 = 0.2670.$$

(b)

$$\begin{aligned}
\mathsf{p}_5 = \binom{7}{5} \left(1 - 0.9^{10}\right)^5 \left(0.9^{10}\right)^2 &= 21 \times 0.6513^5 \times 0.3487^2 = 0.2993. \\
\mathsf{p}_6 = \binom{7}{6} \left(1 - 0.9^{10}\right)^6 \left(0.9^{10}\right)^1 &= 7 \times 0.6513^6 \times 0.3487^1 = 0.1863. \\
\mathsf{p}_7 = \binom{7}{7} \left(1 - 0.9^{10}\right)^7 \left(0.9^{10}\right)^0 &= 0.6513^7 = 0.0497.
\end{aligned}$$

The probability of at least one defective in four or more of the batches is

$$\mathsf{p}_4 + \mathsf{p}_5 + \mathsf{p}_6 + \mathsf{p}_7 = 0.8023.$$

12.

(a) Let Y be the number of companies to which the engineer is called and let A denote the event that the engineer is called to company A.

(i) $\mathsf{P}(Y = 4) = 0.1^4 = 0.0001.$

(ii) $\mathsf{P}(Y \geq 3) = \binom{4}{3} \times 0.1^3 \times 0.9^1 + 0.1^4 = 0.0037.$

(iii) $$\mathsf{P}(Y = 4 \mid Y \geq 1) = \frac{\mathsf{P}(Y = 4 \cap Y \geq 1)}{\mathsf{P}(Y \geq 1)}$$

$$= \frac{\mathsf{P}(Y = 4)}{\mathsf{P}(Y \geq 1)} = \frac{0.0001}{1 - 0.9^4} = \frac{0.0001}{0.3439} = \frac{1}{3439} = 0.0003.$$

(iv) $$\mathsf{P}(Y = 4 \mid A) = \frac{\mathsf{P}(Y = 4 \cap A)}{\mathsf{P}(A)}$$

$$\frac{\mathsf{P}(Y = 4)}{\mathsf{P}(A)} = \frac{0.0001}{0.1} = 0.0010.$$

(b) The mean is $\mathsf{E}(Y) = 4 \times 0.1 = 0.4$. The variance is $V(Y) = 4 \times 0.1 \times 0.9 = 0.36$.

Answers

13. Let D denote the event that the chosen machine is defective and $\bar{D}$ denote the event "not D".
Let Y be the number of imperfect articles in the sample of five.
Then

$$\begin{aligned}
\mathsf{P}(D \mid Y=2) &= \frac{\mathsf{P}(D)\times\mathsf{P}(Y=2\mid D)}{\mathsf{P}(D)\times\mathsf{P}(Y=2\mid D)+\mathsf{P}(\bar{D})\times\mathsf{P}(Y=2\mid \bar{D})}\\
&= \frac{\frac{2}{5}\times\begin{pmatrix}5\\2\end{pmatrix}\times 0.2^2\times 0.8^3}{\frac{2}{5}\times\begin{pmatrix}5\\2\end{pmatrix}\times 0.2^2\times 0.8^3+\frac{3}{5}\times\begin{pmatrix}5\\2\end{pmatrix}\times 0.1^2\times 0.9^3}\\
&= \frac{2\times 0.2^2\times 0.8^3}{2\times 0.2^2\times 0.8^3+3\times 0.1^2\times 0.9^3}\\
&= \frac{0.04096}{0.04096+0.02187}=0.6519.
\end{aligned}$$

14.

(a) (i) $$p_3=\begin{pmatrix}20\\3\end{pmatrix}0.1^3\times 0.9^{17}=\frac{20\times 19\times 18}{1\times 2\times 3}\times 0.1^3\times 0.9^7=0.190.$$

(ii)

$$\begin{aligned}
p_2 &= \begin{pmatrix}20\\2\end{pmatrix}0.1^2\times 0.9^{18}=\frac{3}{18}\times 9\times p_3=0.28518\\
p_1 &= \begin{pmatrix}20\\1\end{pmatrix}0.1\times 0.9^{19}=\frac{2}{19}\times 9\times p_2=0.27017\\
p_0 &= \begin{pmatrix}20\\0\end{pmatrix}0.9^{20}=0.12158.
\end{aligned}$$

The total probability is 0.867.

(iii) The required probability is the probability of at most 2 out of 16.

$$\begin{aligned}
p_0' &= \mathsf{P}(0 \text{ out of } 16)=0.9^{16}=0.185302\\
p_1' &= \mathsf{P}(1 \text{ out of } 16)=\frac{16}{9}\times p_0'=0.3294258\\
p_2' &= \mathsf{P}(2 \text{ out of } 16)=\frac{15}{2}\times\frac{1}{9}\times p_1'=0.2745215
\end{aligned}$$

(b)

$$\frac{0.2\begin{pmatrix}4\\1\end{pmatrix}\times 0.3^1\times 0.7^3}{0.2\begin{pmatrix}4\\1\end{pmatrix}\times 0.3^1\times 0.7^3+0.9\begin{pmatrix}4\\1\end{pmatrix}\times 0.1^1\times 0.9^3} = \frac{0.02058}{0.02058+0.05832}=0.2608.$$

The Poisson Distribution

37.3

Introduction

In this Section we introduce a probability model which can be used when the outcome of an experiment is a random variable taking on positive integer values and where the only information available is a measurement of its average value. This has widespread applications, for example in analysing traffic flow, in fault prediction on electric cables and in the prediction of randomly occurring accidents. We shall look at the Poisson distribution in two distinct ways. Firstly, as a distribution in its own right. This will enable us to apply statistical methods to a set of problems which cannot be solved using the binomial distribution. Secondly, as an approximation to the binomial distribution $X \sim B(n, p)$ in the case where n is large and p is small. You will find that this approximation can often save the need to do much tedious arithmetic.

Prerequisites

Before starting this Section you should ...

- understand the concepts of probability
- understand the concepts and notation for the binomial distribution

Learning Outcomes

On completion you should be able to ...

- recognise and use the formula for probabilities calculated from the Poisson model
- use the recurrence relation to generate a succession of probabilities
- use the Poisson model to obtain approximate values for binomial probabilities

1. The Poisson approximation to the binomial distribution

The probability of the outcome $X = r$ of a set of Bernoulli trials can always be calculated by using the formula

$$\mathrm{P}(X = r) = {}^nC_r q^{n-r} p^r$$

given above. Clearly, for very large values of n the calculation can be rather tedious, this is particularly so when very small values of p are also present. In the situation when n is large and p is small and the product np is constant we can take a different approach to the problem of calculating the probability that $X = r$. In the table below the values of $\mathrm{P}(X = r)$ have been calculated for various combinations of n and p under the constraint that $np = 1$. You should try some of the calculations for yourself using the formula given above for some of the **smaller** values of n.

Probability of X successes

n	p	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$
4	0.25	0.316	0.422	0.211	0.047	0.004		
5	0.20	0.328	0.410	0.205	0.051	0.006	0.000	
10	0.10	0.349	0.387	0.194	0.058	0.011	0.001	0.000
20	0.05	0.359	0.377	0.189	0.060	0.013	0.002	0.000
100	0.01	0.366	0.370	0.185	0.061	0.014	0.003	0.001
1000	0.001	0.368	0.368	0.184	0.061	0.015	0.003	0.001
10000	0.0001	0.368	0.368	0.184	0.061	0.015	0.003	0.001

Each of the binomial distributions given has a mean given by $np = 1$. Notice that the probabilities that $X = 0, 1, 2, 3, 4, \ldots$ approach the values $0.368, 0.368, 0.184, \ldots$ as n increases.

If we have to determine the probabilities of success when large values of n and small values of p are involved it would be very convenient if we could do so without having to construct tables. In fact we can do such calculations by using the Poisson distribution which, under certain constraints, may be considered as an approximation to the binomial distribution.

By considering simplifications applied to the binomial distribution subject to the conditions

1. n is large
2. p is small
3. $np = \lambda$ (λ a constant)

we can derive the formula

$$\mathrm{P}(X = r) = \mathrm{e}^{-\lambda}\frac{\lambda^r}{r!} \quad \text{as an approximation to } \mathrm{P}(X = r) = {}^nC_r q^{n-r} p^r.$$

This is the Poisson distribution given previously. We now show how this is done. We know that the binomial distribution is given by

$$(q + p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{2!}q^{n-2}p^2 + \cdots + \frac{n(n-1)\ldots(n-r+1)}{r!}q^{n-r}p^r + \cdots + p^n$$

Condition (2) tells us that since p is small, $q = 1 - p$ is approximately equal to 1. Applying this to the terms of the binomial expansion above we see that the right-hand side becomes

$$1 + np + \frac{n(n-1)}{2!}p^2 + \cdots + \frac{n(n-1)\ldots(n-r+1)}{r!}p^r + \cdots + p^n$$

Applying condition (1) allows us to approximate terms such as $(n-1),(n-2),\ldots$ to n (mathematically, we are allowing $n \to \infty$) and the right-hand side of our expansion becomes

$$1 + np + \frac{n^2}{2!}p^2 + \cdots + \frac{n^r}{r!}p^r + \ldots$$

Note that the term $p^n \to 0$ under these conditions and hence has been omitted.
We now have the series

$$1 + np + \frac{(np)^2}{2!} + \cdots + \frac{(np)^r}{r!} + \ldots$$

which, using condition (3) may be written as

$$1 + \lambda + \frac{(\lambda)^2}{2!} + \cdots + \frac{(\lambda)^r}{r!} + \ldots$$

You may recognise this as the expansion of e^{λ}.

If we are to be able to claim that the terms of this expansion represent probabilities, we must be sure that the sum of the terms is 1. We divide by e^{λ} to satisfy this condition. This gives the result

$$\frac{\mathrm{e}^{\lambda}}{\mathrm{e}^{\lambda}} = 1 = \frac{1}{\mathrm{e}^{\lambda}}\left(1 + \lambda + \frac{(\lambda)^2}{2!} + \cdots + \frac{(\lambda)^r}{r!} + \ldots\right)$$
$$= \mathrm{e}^{-\lambda} + \mathrm{e}^{-\lambda}\lambda + \mathrm{e}^{-\lambda}\frac{\lambda^2}{2!} + \mathrm{e}^{-\lambda}\frac{\lambda^3}{3!} + \cdots + \mathrm{e}^{-\lambda}\frac{\lambda^r}{r!} + \cdots +$$

The terms of this expansion are very good approximations to the corresponding binomial expansion under the conditions

1. n is large
2. p is small
3. $np = \lambda$ (λ constant)

The Poisson approximation to the binomial distribution is summarized below.

Key Point 6

Poisson Approximation to the Binomial Distribution

Assuming that n is large, p is small and that np is constant, the terms

$$\mathrm{P}(X = r) = {}^nC_r(1-p)^{n-r}p^r$$

of a binomial distribution may be closely approximated by the terms

$$\mathrm{P}(X = r) = \mathrm{e}^{-\lambda}\frac{\lambda^r}{r!}$$

of the Poisson distribution for corresponding values of r.

Example 12

We introduced the binomial distribution by considering the following scenario. A worn machine is known to produce 10% defective components. If the random variable X is the number of defective components produced in a run of 3 components, find the probabilities that X takes the values 0 to 3.

Suppose now that a similar machine which is known to produce 1% defective components is used for a production run of 40 components. We wish to calculate the probability that two defective items are produced. Essentially we are assuming that $X \sim B(40, 0.01)$ and are asking for $\mathsf{P}(X = 2)$. We use both the binomial distribution and its Poisson approximation for comparison.

Solution

Using the binomial distribution we have the solution

$$\mathsf{P}(X = 2) = \ {}^{40}C_2(0.99)^{40-2}(0.01)^2 = \frac{40 \times 39}{1 \times 2} \times 0.99^{38} \times 0.01^2 = 0.0532$$

Note that the arithmetic involved is unwieldy. Using the Poisson approximation we have the solution

$$\mathsf{P}(X = 2) = \mathsf{e}^{-0.4}\frac{0.4^2}{2!} = 0.0536$$

Note that the arithmetic involved is simpler and the approximation is reasonable.

Practical considerations

In practice, we can use the Poisson distribution to very closely approximate the binomial distribution provided that the product np is constant with

$n \geq 100$ and $p \leq 0.05$

Note that this is not a hard-and-fast rule and we simply say that

'the larger n is the better and the smaller p is the better provided that np is a sensible size.'

The approximation remains good provided that $np < 5$ for values of n as low as 20.

Task

Mass-produced needles are packed in boxes of 1000. It is believed that 1 needle in 2000 on average is substandard. What is the probability that a box contains 2 or more defectives? The correct model is the binomial distribution with $n = 1000,\ p = \frac{1}{2000}$ (and $q = \frac{1999}{2000}$).

(a) Using the binomial distribution calculate $\mathsf{P}(X=0),\ \mathsf{P}(X=1)$ and hence $\mathsf{P}(X \geq 2)$:

Your solution

Answer

$$\mathsf{P}(X=0) = \left(\frac{1999}{2000}\right)^{1000} = 0.60645$$

$$\mathsf{P}(X=1) = 1000\left(\frac{1999}{2000}\right)^{999} \times \left(\frac{1}{2000}\right) = \frac{1}{2}\left(\frac{1999}{2000}\right)^{999} = 0.30338$$

$\therefore \quad \mathsf{P}(X=0) + \mathsf{P}(X=1) = 0.60645 + 0.30338 = 0.90983 \simeq 0.9098$ (4 d.p.)

Hence P(2 or more defectives) $\simeq 1 - 0.9098 = 0.0902$.

(b) Now choose a suitable value for λ in order to use a Poisson model to approximate the probabilities:

Your solution

$\lambda =$

Answer

$\lambda = np = 1000 \times \dfrac{1}{2000} = \frac{1}{2}$

Now recalculate the probability that there are 2 or more defectives using the Poisson distribution with $\lambda = \frac{1}{2}$:

Your solution

$\mathsf{P}(X=0) =$

$\mathsf{P}(X=1) =$

$\therefore \quad$ P(2 or more defectives)$=$

Answer

$\mathsf{P}(X=0) = e^{-\frac{1}{2}}, \quad \mathsf{P}(X=1) = \frac{1}{2}e^{-\frac{1}{2}}$

$\therefore \quad \mathsf{P}(X=0) + \mathsf{P}(X=1) = \frac{3}{2}\, e^{-\frac{1}{2}} = 0.9098$ (4 d.p.)

Hence P(2 or more defectives) $\simeq 1 - 0.9098 = 0.0902$.

In the above Task we have obtained the same answer to 4 d.p., as the exact binomial calculation, essentially because p was so small. We shall not always be so lucky!

Example 13

In the manufacture of glassware, bubbles can occur in the glass which reduces the status of the glassware to that of a 'second'. If, on average, one in every 1000 items produced has a bubble, calculate the probability that exactly six items in a batch of three thousand are seconds.

Solution

Suppose that X = number of items with bubbles, then $X \sim B(3000, 0.001)$

Since $n = 3000 > 100$ and $p = 0.001 < 0.005$ we can use the Poisson distribution with $\lambda = np = 3000 \times 0.001 = 3$. The calculation is:

$$\mathrm{P}(X = 6) = \mathrm{e}^{-3}\frac{3^6}{6!} \approx 0.0498 \times 1.0125 \approx 0.05$$

The result means that we have about a 5% chance of finding exactly six seconds in a batch of three thousand items of glassware.

Example 14

A manufacturer produces light-bulbs that are packed into boxes of 100. If quality control studies indicate that 0.5% of the light-bulbs produced are defective, what percentage of the boxes will contain:

(a) no defective? (b) 2 or more defectives?

Solution

As n is large and p, the P(defective bulb), is small, use the Poisson approximation to the binomial probability distribution. If X = number of defective bulbs in a box, then

$$X \sim \mathrm{P}(\mu) \text{ where } \mu = n \times p = 100 \times 0.005 = 0.5$$

(a) $\mathrm{P}(X = 0) = \dfrac{e^{-0.5}(0.5)^0}{0!} = \dfrac{e^{-0.5}(1)}{1} = 0.6065 \approx 61\%$

(b) $\mathrm{P}(X = 2 \text{ or more}) = \mathrm{P}(X = 2) + \mathrm{P}(X = 3) + \mathrm{P}(X = 4) + \ldots$ but it is easier to consider:

$$\mathrm{P}(X \geq 2) = 1 - [\mathrm{P}(X = 0) + \mathrm{P}(X = 1)]$$

$$\mathrm{P}(X = 1) = \frac{e^{-0.5}(0.5)^1}{1!} = \frac{e^{-0.5}(0.5)}{1} = 0.3033$$

i.e. $\mathrm{P}(X \geq 2) = 1 - [0.6065 + 0.3033] = 0.0902 \approx 9\%$

2. The Poisson distribution

The Poisson distribution is a probability model which can be used to find the probability of a single event occurring a given number of times in an interval of (usually) time. The occurrence of these events must be determined by chance alone which implies that information about the occurrence of any one event cannot be used to predict the occurrence of any other event. It is worth noting that only the *occurrence* of an event can be counted; the *non-occurrence* of an event cannot be counted. This contrasts with Bernoulli trials where we know the number of trials, the number of events occurring and therefore the number of events not occurring.

The Poisson distribution has widespread applications in areas such as analysing traffic flow, fault prediction in electric cables, defects occurring in manufactured objects such as castings, email messages arriving at a computer and in the prediction of randomly occurring events or accidents. One well known series of accidental events concerns Prussian cavalry who were killed by horse kicks. Although not discussed here (death by horse kick is hardly an engineering application of statistics!) you will find accounts in many statistical texts. One example of the use of a Poisson distribution where the events are not necessarily time related is in the prediction of fault occurrence along a long weld - faults may occur anywhere along the length of the weld. A similar argument applies when scanning castings for faults - we are looking for faults occurring in a volume of material, not over an interval if time.

The following definition gives a theoretical underpinning to the Poisson distribution.

Definition of a Poisson process

Suppose that events occur at random throughout an interval. Suppose further that the interval can be divided into subintervals which are so small that:

1. the probability of more than one event occurring in the subinterval is zero
2. the probability of one event occurring in a subinterval is proportional to the length of the subinterval
3. an event occurring in any given subinterval is independent of any other subinterval

then the random experiment is known as a **Poisson process**.

The word 'process' is used to suggest that the experiment takes place over time, which is the usual case. If the average number of events occurring in the interval (not subinterval) is $\lambda\ (> 0)$ then the random variable X representing the actual number of events occurring in the interval is said to have a Poisson distribution and it can be shown (we omit the derivation) that

$$\mathsf{P}(X = r) = \mathsf{e}^{-\lambda}\frac{\lambda^r}{r!} \qquad r = 0, 1, 2, 3, \ldots$$

The following Key Point provides a summary.

Key Point 7

The Poisson Probabilities

If X is the random variable

'number of occurrences in a given interval'

for which the average rate of occurrence is λ then, according to the **Poisson** model, the probability of r occurrences in that interval is given by

$$\mathsf{P}(X = r) = e^{-\lambda}\,\frac{\lambda^r}{r!} \qquad r = 0, 1, 2, 3, \ldots$$

Task

Using the Poisson distribution $\mathsf{P}(X = r) = \mathrm{e}^{-\lambda}\frac{\lambda^r}{r!}$ write down the formulae for $\mathsf{P}(X = 0)$, $\mathsf{P}(X = 1)$, $\mathsf{P}(X = 2)$ and $\mathsf{P}(X = 6)$, noting that $0! = 1$.

Your solution

$\mathsf{P}(X = 0) =$

$\mathsf{P}(X = 1) =$

$\mathsf{P}(X = 2) =$

$\mathsf{P}(X = 6) =$

Answer

$$\mathsf{P}(X = 0) = e^{-\lambda} \times \frac{\lambda^0}{0!} = e^{-\lambda} \times \frac{1}{1} \equiv e^{-\lambda} \qquad \mathsf{P}(X = 1) = e^{-\lambda} \times \frac{\lambda}{1!} = \lambda e^{-\lambda}$$

$$\mathsf{P}(X = 2) = e^{-\lambda} \times \frac{\lambda^2}{2!} = \frac{\lambda^2}{2} e^{-\lambda} \qquad \mathsf{P}(X = 6) = e^{-\lambda} \times \frac{\lambda^6}{6!} = \frac{\lambda^6}{720} e^{-\lambda}$$

Task

Calculate $\mathsf{P}(X = 0)$ to $\mathsf{P}(X = 5)$ when $\lambda = 2$, accurate to 4 d.p.

Your solution

Answer

r	0	1	2	3	4	5
$\mathsf{P}(X = r)$	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361

Notice how the values for $\mathsf{P}(X = r)$ in the above answer increase, stay the same and then decrease relatively rapidly (due to the significant increase in $r!$ with increasing r). Here two of the probabilities are equal and this will always be the case when λ is an integer.

In this last Task we only went up to $\mathsf{P}(X = 5)$ and calculated each entry separately. However, each probability need not be calculated directly. We can use the following relations (which can be checked from the formulae for $\mathsf{P}(X = r)$) to get the next probability from the previous one:

$$\mathsf{P}(X = 1) = \frac{\lambda}{1}\,\mathsf{P}(X = 0)\ , \quad \mathsf{P}(X = 2) = \frac{\lambda}{2}\,\mathsf{P}(X = 1), \qquad \mathsf{P}(X = 3) = \frac{\lambda}{3}\,\mathsf{P}(X = 2)\ ,\ \text{etc.}$$

Key Point 8

Recurrence Relation for Poisson Probabilities

In general, for ease of calculation the **recurrence relation** below can be used

$$\mathsf{P}(X = r) = \frac{\lambda}{r}\,\mathsf{P}(X = r - 1) \qquad \text{for } r \geq 1.$$

Example 15

Calculate the value for $\mathsf{P}(X = 6)$ to extend the Table in the previous Task using the recurrence relation and the value for $\mathsf{P}(X = 5)$.

Solution

The recurrence relation gives the formula

$$\mathsf{P}(X = 6) = \frac{2}{6}\mathsf{P}(X = 5) = \frac{1}{3} \times 0.0361 = 0.0120$$

We now look further at the Poisson distribution by considering an example based on traffic flow.

Example 16

Suppose it has been observed that, on average, 180 cars per hour pass a specified point on a particular road in the morning rush hour. Due to impending roadworks it is estimated that congestion will occur closer to the city centre if more than 5 cars pass the point in any one minute. What is the probability of congestion occurring?

Solution

We note that we cannot use the binomial model since we have no values of n and p. Essentially we are saying that there is no fixed number (n) of cars passing the specified point and that we have no way of estimating p. The only information available is the average rate at which cars pass the specified point.

Let X be the random variable $X =$ number of cars arriving in any minute. We need to calculate the probability that more than 5 cars arrive in any one minute. Note that in order to do this we need to convert the information given on the average rate (cars arriving per hour) into a value for λ (cars arriving per minute). This gives the value $\lambda = 3$.

Using $\lambda = 3$ to calculate the required probabilities gives:

r	0	1	2	3	4	5	Sum
$\mathsf{P}(X = r)$	0.04979	0.149361	0.22404	0.22404	0.168031	0.10082	0.91608

To calculate the required probability we note that

P(more than 5 cars arrive in one minute) $= 1 -$ P(5 cars or less arrive in one minute)

Thus

$$\begin{aligned}\mathsf{P}(X > 5) &= 1 - \mathsf{P}(X \le 5)\\ &= 1 - \mathsf{P}(X = 0) - \mathsf{P}(X = 1) - \mathsf{P}(X = 2) - \mathsf{P}(X = 3) - \mathsf{P}(X = 4) - \mathsf{P}(X = 5)\end{aligned}$$

Then P(more than 5) $= 1 - 0.91608 = 0.08392 = 0.0839$ (4 d.p).

Example 17

The mean number of bacteria per millilitre of a liquid is known to be 6. Find the probability that in 1 ml of the liquid, there will be:

(a) 0, (b) 1, (c) 2, (d) 3, (e) less than 4, (f) 6 bacteria.

Solution

Here we have an *average rate of occurrences* but no estimate of the probability so it looks as though we have a Poisson distribution with $\lambda = 6$. Using the formula in Key Point 7 we have:

(a) $\mathsf{P}(X = 0) = e^{-6}\dfrac{6^0}{0!} = 0.00248.$

That is, the probability of having no bacteria in 1 ml of liquid is 0.00248

(b) $\mathsf{P}(X = 1) = \dfrac{\lambda}{1} \times \mathsf{P}(X = 0) = 6 \times 0.00248 = 0.0149.$

That is, the probability of having 1 bacteria in 1 ml of liquid is 0.0149

(c) $\mathsf{P}(X = 2) = \dfrac{\lambda}{2} \times \mathsf{P}(X = 1) = \dfrac{6}{2} \times 0.01487 = 0.0446.$

That is, the probability of having 2 bacteria in 1 ml of liquid is 0.0446

(d) $\mathsf{P}(X = 3) = \dfrac{\lambda}{3} \times \mathsf{P}(X = 2) = \dfrac{6}{3} \times 0.04462 = 0.0892.$

That is, the probability of having 3 bacteria in 1 ml of liquid is 0.0892

(e) $\mathsf{P}(X < 4) = \mathsf{P}(X = 0) + \mathsf{P}(X = 1) + \mathsf{P}(X = 2) + \mathsf{P}(X = 3) = 0.1512$

(f) $\mathsf{P}(X = 6) = e^{-6}\dfrac{6^6}{6!} = 0.1606$

Note that in working out the first 6 answers, which link together, all the digits were kept in the calculator to ensure accuracy. Answers were rounded off only when written down.

Never copy down answers correct to, say, 4 decimal places and then use those rounded figures to calculate the next figure as rounding-off errors will become greater at each stage. If you did so here you would get answers 0.0025, 0.0150, 0.0450, 0.9000 and $\mathsf{P}(X < 4) = 0.1525$. The difference is not great but could be significant.

Task

A Council is considering whether to base a recovery vehicle on a stretch of road to help clear incidents as quickly as possible. The road concerned carries over 5000 vehicles during the peak rush hour period. Records show that, on average, the number of incidents during the morning rush hour is 5. The Council won't base a vehicle on the road if the probability of having more than 5 incidents in any one morning is less than 30%. Based on this information should the Council provide a vehicle?

Your solution

(Do the calculation on separate paper and record the main results here.)

Answer

We need to calculate the probability that more than 5 incidents occur i.e. $\mathsf{P}(X > 5)$. To find this we use the fact that $\mathsf{P}(X > 5) = 1 - \mathsf{P}(X \leq 5)$. Now, for this problem:

$$\mathsf{P}(X = r) = e^{-5}\frac{5^r}{r!}$$

Writing answers to 5 d.p. gives:

$$\begin{aligned} \mathsf{P}(X = 0) &= e^{-5}\frac{5^0}{0!} = 0.00674 \\ \mathsf{P}(X = 1) &= 5 \times \mathsf{P}(X = 0) = 0.03369 \\ \mathsf{P}(X = 2) &= \frac{5}{2} \times \mathsf{P}(X = 1) = 0.08422 \\ \mathsf{P}(X = 3) &= \frac{5}{3} \times \mathsf{P}(X = 2) = 0.14037 \\ \mathsf{P}(X = 4) &= \frac{5}{4} \times \mathsf{P}(X = 3) = 0.17547 \\ \mathsf{P}(X = 5) &= \frac{5}{5} \times \mathsf{P}(X = 4) = 0.17547 \end{aligned}$$

$$\begin{aligned} \mathsf{P}(X \leq 5) &= \mathsf{P}(X = 0) + \mathsf{P}(X = 1) + \mathsf{P}(X = 2) + \mathsf{P}(X = 3) + \mathsf{P}(X = 4) + \mathsf{P}(X = 5) \\ &= 0.61596 \end{aligned}$$

The probability of more than 5 incidents is $\mathsf{P}(X > 5) = 1 - \mathsf{P}(X \leq 5) = 0.38403$, which is 38.4% (to 3 s.f.) so the Council should provide a vehicle.

3. Expectation and variance of the poisson distribution

The expectation and variance of the Poisson distribution can be derived directly from the definitions which apply to any discrete probability distribution. However, the algebra involved is a little lengthy. Instead we derive them from the binomial distribution from which the Poisson distribution is derived.

Intuitive Explanation

One way of deriving the mean and variance of the Poisson distribution is to consider the behaviour of the binomial distribution under the following conditions:

1. n is large 2. p is small 3. $np = \lambda$ (a constant)

Recalling that the expectation and variance of the binomial distribution are given by the results

$$\mathsf{E}(X) = np \qquad \text{and} \qquad \mathsf{V}(X) = np(1-p) = npq$$

it is reasonable to assert that condition (2) implies, since $q = 1 - p$, that q is approximately 1 and so the expectation and variance are given by

$$\mathsf{E}(X) = np \qquad \text{and} \qquad \mathsf{V}(X) = npq \approx np$$

In fact the algebraic derivation of the expectation and variance of the Poisson distribution shows that these results are in fact *exact*.

Note that the expectation and the variance are equal.

Key Point 9

The Poisson Distribution

If X is the random variable $\{$number of occurrences in a given interval$\}$

for which the average rate of occurrences is λ and X can assume the values $0, 1, 2, 3, \ldots$ and the probability of r occurrences in that interval is given by

$$\mathsf{P}(X = r) = \mathsf{e}^{-\lambda}\frac{\lambda^r}{r!}$$

then the expectation and variance of the distribution are given by the formulae

$$\mathsf{E}(X) = \lambda \qquad \text{and} \qquad \mathsf{V}(X) = \lambda$$

For a Poisson distribution the Expectation and Variance are equal.

Exercises

1. Large sheets of metal have faults in random positions but on average have 1 fault per 10 m^2.

   What is the probability that a sheet 5 m $\times$ 8 m will have at most one fault?

2. If 250 litres of water are known to be polluted with 10^6 bacteria what is the probability that a sample of 1 cc of the water contains no bacteria?

3. Suppose vehicles arrive at a signalised road intersection at an average rate of 360 per hour and the cycle of the traffic lights is set at 40 seconds. In what percentage of cycles will the number of vehicles arriving be (a) exactly 5, (b) less than 5? If, after the lights change to green, there is time to clear only 5 vehicles before the signal changes to red again, what is the probability that waiting vehicles are not cleared in one cycle?

4. Previous results indicate that 1 in 1000 transistors are defective on average.

   (a) Find the probability that there are 4 defective transistors in a batch of 2000.

   (b) What is the largest number, N, of transistors that can be put in a box so that the probability of no defectives is at least $1/2$?

5. A manufacturer sells a certain article in batches of 5000. By agreement with a customer the following method of inspection is adopted: A sample of 100 items is drawn at random from each batch and inspected. If the sample contains 4 or fewer defective items, then the batch is accepted by the customer. If more than 4 defectives are found, every item in the batch is inspected. If inspection costs are 75 p per hundred articles, and the manufacturer normally produces 2% of defective articles, find the average inspection costs per batch.

6. A book containing 150 pages has 100 misprints. Find the probability that a particular page contains (a) no misprints, (b) 5 misprints, (c) at least 2 misprints, (d) more than 1 misprint.

7. For a particular machine, the probability that it will break down within a week is 0.009. The manufacturer has installed 800 machines over a wide area. Calculate the probability that (a) 5, (b) 9, (c) less than 5, (d) more than 4 machines breakdown in a week.

8. At a given university, the probability that a member of staff is absent on any one day is 0.001. If there are 800 members of staff, calculate the probabilities that the number absent on any one day is (a) 6, (b) 4, (c) 2, (d) 0, (e) less than 3, (f) more than 1.

9. The number of failures occurring in a machine of a certain type in a year has a Poisson distribution with mean 0.4. In a factory there are ten of these machines. What is

   (a) the expected total number of failures in the factory in a year?

   (b) the probability that there are fewer than two failures in the factory in a year?

Exercises continued

10. A factory uses tools of a particular type. From time to time failures in these tools occur and they need to be replaced. The number of such failures in a day has a Poisson distribution with mean 1.25. At the beginning of a particular day there are five replacement tools in stock. A new delivery of replacements will arrive after four days. If all five spares are used before the new delivery arrives then further replacements cannot be made until the delivery arrives. Find

    (a) the probability that three replacements are required over the next four days.

    (b) the expected number of replacements actually made over the next four days.

Answers

1. Poisson Process. In a sheet size 40 m^2 **we expect** 4 faults

$$\therefore \qquad \lambda = 4 \qquad \mathsf{P}(X = r) = \lambda^r e^{-\lambda}/r!$$

$$\mathsf{P}(X \leq 1) = \mathsf{P}(X = 0) + \mathsf{P}(X = 1) = e^{-4} + 4e^{-4} = 0.0916$$

2. In 1 cc **we expect** 4 bacteria$(= 10^6/250000) \quad \therefore \qquad \lambda = 4$

$$\mathsf{P}(X = 0) = e^{-4} = 0.0183$$

3. In 40 seconds **we expect** 4 vehicles $\quad \therefore \qquad \lambda = 4$

   (a) P (exactly 5) $= \lambda^5 e^{-\lambda}/5! = 0.15629$ i.e. in 15.6% of cycles

   (b) P (less than 5) $= e^{-\lambda}\left[1 + \lambda^4 + \dfrac{\lambda^2}{2!} + \dfrac{\lambda^3}{3!} + \dfrac{\lambda^4}{4!}\right]$

$$= e^{-4}\left[1 + 4 + 8 + \frac{32}{3} + \frac{32}{3}\right] = 0.6288$$

   Vehicles will not be cleared if more than 5 are waiting.

   P (greater than 5) $= 1 -$ P (exactly 5) $-$P (less than 5)

   $= 1 - 0.15629 - 0.6288 = 0.2148$

4 (a) Poisson approximation to binomial

$$\lambda = np = 2000.\frac{1}{1000} = 2$$

$$\mathsf{P}(X = 4) = \lambda^4 e^{-\lambda}/4! = 16e^{-2}/24 = 0.09022$$

(b) $\lambda = Np = N/1000; \quad \mathsf{P}(X = 0) = \dfrac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda} = e^{-N/1000}$

$$e^{-N/1000} = 0.5 \qquad \therefore \qquad \frac{-N}{1000} = \ln(0.5)$$

$\therefore \qquad N = 693.147 \qquad$ choose $N = 693$ or less.

Answers

5. P(defective) $= 0.02$. Poisson approximation to binomial $\lambda = np = 100(0.02) = 2$

P(4 or fewer defectives in sample of 100)

$$= \text{P}(X = 0) + \text{P}(X = 1) + \text{P}(X = 2) + \text{P}(X = 3) + \text{P}(X = 4)$$

$$= e^{-2} + 2e^{-2} + \frac{2^2}{2}e^{-2} + \frac{2^3}{3!}e^{-2} + \frac{2^4}{4!}e^{-2} = 0.947347$$

Inspection costs

Cost c	75	75×50
P$(X = c)$	0.947347	0.0526

E(Cost) $= 75(0.947347) + 75 \times 50(0.0526) = 268.5$ p

6. (a) 0.51342 (b) 0.00056, (c) 0.14430, (d) 0.14430

7. (a) 0.12038, (b) 0.10698, (c) 0.15552, (d) 0.84448

8. (a) 0.00016, (b) 0.00767, (c) 0.14379, (d) 0.44933, (e) 0.95258, (f) 0.19121

9. Let X be the total number of failures.

(a) E$(X) = 10 \times 0.4 = 4.$

(b) P$(X < 2) = $ P$(X = 0) + $ P$(X = 1) = e^{-4} + 4e^{-4} = 5e^{-4} = 0.0916.$

10. Let the number required over 4 days be X. Then E$(X) = 4 \times 1.25 = 5$ and $X \sim$ Poisson(5).

(a) P$(X = 3) = \dfrac{e^{-5}5^3}{3!} = 0.1404.$

(b) Let R be the number of replacements made.

$$\text{E}(R) = 0 \times \text{P}(X = 0) + \cdots + 4 \times \text{P}(X = 4) + 5 \times \text{P}(X \geq 5),$$

and

$$\text{P}(X \geq 5) = 1 - [\text{P}(X = 0) + \cdots + \text{P}(X = 4)]$$

so

$$\begin{aligned} \text{E}(R) &= 5 - 5 \times \text{P}(X = 0) - \cdots - 1 \times \text{P}(X = 4) \\ &= 5 - e^{-5}\left[5 \times \frac{5^0}{0!} + 4 \times \frac{5^1}{1!} + \cdots + 1 \times \frac{5^4}{4!}\right] \\ &= 5 - 0.8773 \\ &= 4.123. \end{aligned}$$

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. 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 ...

- understand the concepts of probability
- understand the notation ${}^nC_r$ used in probability calculations

Learning Outcomes

On completion 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:

$$\mathsf{P}(X = r) = {}^nC_r(1-p)^{n-r}p^r$$

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

Example 18

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:

(a) the first ring is defective;

(b) the second ring is defective given that the first one is defective.

Solution

(a) The probability that the first ring is defective is clearly $\dfrac{10}{100} = \dfrac{1}{10}$.

(b) 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 10

Hypergeometric Distribution

The distribution given by

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

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 19

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

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

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 11

Expectation and Variance of the Hypergeometric Distribution

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

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

are given by

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

Example 20

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

Solution

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

Hence

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

and

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

Task

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

Answer

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

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

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

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

Task

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.

(a) 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.

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

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

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

Your solution

Answer

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

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

Hence

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

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

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

(b) (i) Case $D = 5$

P(Accept batch with 5 defectives) is

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

(b) (ii) Case $D = 3$

P(Reject batch with 3 defectives) is

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

Exercise

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 $d = 1$.

Answer

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

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

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

$$\mathsf{P}(X=0)+\mathsf{P}(X=1)=\frac{\begin{pmatrix}21\\10\end{pmatrix}\begin{pmatrix}9\\0\end{pmatrix}}{\begin{pmatrix}30\\10\end{pmatrix}}+\frac{\begin{pmatrix}21\\9\end{pmatrix}\begin{pmatrix}9\\1\end{pmatrix}}{\begin{pmatrix}30\\10\end{pmatrix}}$$

Now

$$\begin{aligned} \begin{pmatrix}21\\10\end{pmatrix}\begin{pmatrix}9\\0\end{pmatrix} &= \frac{21}{10}\times\frac{20}{9}\times\frac{19}{8}\times\frac{18}{7}\times\frac{17}{6}\times\frac{16}{5}\times\frac{15}{4}\times\frac{14}{3}\times\frac{13}{2}\times\frac{12}{1} \\ &= 352716 \end{aligned}$$

$$\begin{aligned} \begin{pmatrix}21\\9\end{pmatrix}\begin{pmatrix}9\\1\end{pmatrix} &= \frac{21}{9}\times\frac{20}{8}\times\frac{19}{7}\times\frac{18}{6}\times\frac{17}{5}\times\frac{16}{4}\times\frac{15}{3}\times\frac{14}{2}\times\frac{13}{1}\times\frac{9}{1} \\ &= 2645370 \end{aligned}$$

$$\begin{aligned} \begin{pmatrix}30\\10\end{pmatrix} &= \frac{30}{10}\times\frac{29}{9}\times\frac{28}{8}\times\frac{27}{7}\times\frac{26}{6}\times\frac{25}{5}\times\frac{24}{4}\times\frac{23}{3}\times\frac{22}{2}\times\frac{21}{1} \\ &= 30045015 \end{aligned}$$

So the probability of acceptance is

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

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