"JUST THE MATHS"

UNIT NUMBER

19.6

PROBABILITY 6
(Statistics for the binomial distribution)

by

A.J.Hobson

UNIT 19.6 - PROBABILITY 6

STATISTICS FOR THE BINOMIAL DISTRIBUTION

19.6.1 CONSTRUCTION OF HISTOGRAMS

Elementary discussion on the presentation of data, in the form of frequency tables, histograms etc., usually involves experiments which are actually carried out.

But we illustrate now how the binomial distribution may be used to estimate the results of a certain kind of experiment before it is performed.

EXAMPLE

For four coins, tossed 32 times, construct a histogram showing the expected number of occurrences of 0,1,2,3,4..... heads.

Solution

Firstly, in a single toss of the four coins, the probability of head (or tail) for each coin is $\frac{1}{2}$.

The terms in the expansion of $\left(\frac{1}{2}+\frac{1}{2}\right)^4$ give the probabilities of exactly 0,1,2,3 and 4 heads, respectively.

The expansion is

$$\left(\frac{1}{2}+\frac{1}{2}\right)^4 \equiv \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + 6\left(\frac{1}{2}\right)^2 + 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4.$$

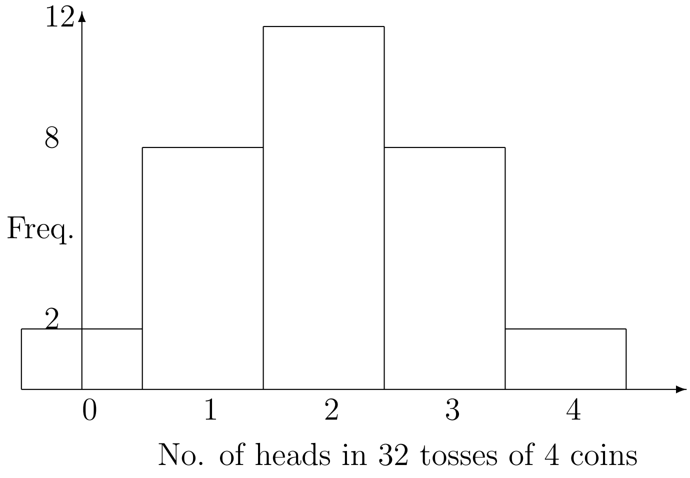
That is,

$$\left(\frac{1}{2}+\frac{1}{2}\right)^4 \equiv \left(\frac{1}{2}\right)^4 (1+4+6+4+1),$$

showing that the probabilities of 0,1,2,3 and 4 heads in a single toss of four coins are $\frac{1}{16}$, $\frac{1}{4}$, $\frac{6}{16}$, $\frac{1}{4}$, and $\frac{1}{16}$, respectively.

Therefore, in 32 tosses of four coins, we may expect 0 heads, twice; 1 head, 8 times; 2 heads, 12 times; 3 heads, 8 times and 4 heads, twice.

The following histogram uses class-intervals for which each member is, in fact, situated at the mid-point:



Notes:

(i) The only reason that the above histogram is symmetrical in shape is that the probability of success and failure are equal to each other, so that the terms of the binomial expansion are, themselves, symmetrical.

(ii) Since the widths of the class-intervals in the above histogram are 1, the areas of the rectangles are equal to their heights. Thus, for example, the total area of the first three rectangles represents the expected number of times of obtaining at most 2 heads in 32 tosses of 4 coins.

19.6.2 MEAN AND STANDARD DEVIATION OF A BINOMIAL DISTRIBUTION

THEOREM

If p is the probability of success of an event in a single trial and q is the probability of its failure, then the binomial distribution, giving the expected frequencies of 0,1,2,3,.....n successes in n trials, has a mean of np and a standard deviation of $\sqrt{npq}$, irrespective of the number of times the experiment is to be carried out.

Proof:

(a) The Mean

From the binomial expansion formula,

$$(q+p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{2!}q^{n-2}p^2 + \frac{n(n-1)(n-2)}{3!}q^{n-3}p^3 + \ldots + nqp^{n-1} + p^n.$$

Hence, if the n trials are made N times, the average number of successes is equal to the following expression, multiplied by N, then divided by N:

$$0 \times q^n + 1 \times nq^{n-1}p + 2 \times \frac{n(n-1)}{2!}q^{n-2}p^2 +$$

$$3 \times \frac{n(n-1)(n-2)}{3!}q^{n-3}p^3 + \;\; ..(n-1) \times nqp^{n-1} + np^n.$$

That is, the mean is

$$np\left(q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2}q^{n-3}p^2 + \;\; \ldots + (n-1)qp^{n-2} + p^{n-1}\right)$$

$$= np(q+p)^{n-1} = np \;\; \text{since} \;\; q+p=1.$$

(b) The Standard Deviation

For the standard deviation, we observe that, if f_r is the frequency of r successes when the n trials are conducted N times, then

$$f_r = N\frac{n!}{(n-r)!r!}q^{n-r}p^r.$$

We use this, first, to establish a result for

$$\sum_{r=0}^{n} r^2 f_r.$$

For example,

$0^2 f_0 = 0.Nq^n = 0 = 0.f_0$.

$1^2 f_1 = 1.Nnq^{n-1}p = 1.f_1$.

$2^2 f_2 =$

$$2Nn(n-1)q^{n-2}p^2 = Nn(n-1)q^{n-2}p^2 + Nn(n-1)p^2q^{n-2}$$

$$= 2f_2 + Nn(n-1)p^2q^{n-2};$$

$3^2 f_3 =$

$$3N\frac{n(n-1)(n-2)}{2!}q^{n-3}p^3 = N\frac{n(n-1)(n-2)}{2!}q^{n-3}p^3 + Nn(n-1)p^2(n-2)q^{n-3}p$$

$$= 3f_3 + Nn(n-1)p^2(n-2)q^{n-3}p;$$

$4^2 f_4 =$

$$4N\frac{n(n-1)(n-2)(n-3)}{3!}q^{n-4}p^4 = N\frac{n(n-1)(n-2)(n-3)}{3!}q^{n-4}p^4 + Nn(n-1)p^2\frac{(n-2)(n-3)}{2!}q^{n-4}p^2$$

$$= 4f_4 + Nn(n-1)p^2\frac{(n-2)(n-3)}{2!}q^{n-4}p^2;$$

and, in general, when $r \geq 2$,

$r^2 f_r =$

$$N\frac{n(n-1)(n-2)....(n-r+1)}{(r-1)!} + Nn(n-1)p^2q^{n-r}p^r = rf_r + Nn(n-1)p^2\frac{(n-2)!}{(n-r)!(r-2)!}q^{n-r}p^{r-2}.$$

This result, together with those for $0^2.f_0$ and $1^2 f_1$, shows that

$$\sum_{r=0}^{n} r^2 f_r = \sum_{r=0}^{n} rf_r + Nn(n-1)p^2 \sum_{r=2}^{n} \frac{(n-2)!}{(n-r)!(r-2)!}q^{n-r}p^{r-2}.$$

That is,

$$\sum_{r=0}^{n} r^2 f_r = Nnp + Nn(n-1)p^2(q+p)^{n-2} = Nnp + Nn(n-1)p^2,$$

since $q + p = 1$.

It was also established, in Unit 18.3, that the standard deviation of a set, $x_1, x_2, x_3, \ldots x_m$, of m observations, with a mean value of $\overline{x}$, is given by the formula

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2 - \overline{x}^2,$$

which, in the present case, may be written

$$\sigma^2 = \frac{1}{N}\sum_{r=0}^{n} r^2 f_r - \frac{1}{N^2}\left(\sum_{r=0}^{n} r f_r\right)^2.$$

Hence,

$$\sigma^2 = \frac{1}{N}\left(Nnp + Nn(n-1)p^2\right) - \frac{1}{N^2}(Nnp)^2,$$

which gives

$$\sigma^2 = np + n^2p^2 - np^2 - n^2p^2 = np(1-p) = npq;$$

and so,

$$\sigma = \sqrt{npq}.$$

ILLUSTRATION

For direct calculation of the mean and the standard deviation for the data in the previous coin-tossing problem, we may use the following table, in which x_i denotes numbers of heads and f_i denotes the corresponding expected frequencies:

x_i	f_i	$f_i x_i$	$f_i x_i^2$
0	2	0	0
1	8	8	8
2	12	24	48
3	8	24	72
4	2	8	32
Totals	32	64	160

The mean is given by

$$\overline{x} = \frac{64}{32} = 2 \quad \text{(obviously)},$$

which agrees with $np = 4 \times \frac{1}{2}$.

The standard deviation is given by

$$\sigma = \sqrt{\left[\frac{160}{32} - 2^2\right]} = 1,$$

which agrees with $\sqrt{npq} = \sqrt{4 \times \frac{1}{2} \times \frac{1}{2}}$.

Note:
If the experiment were carried out N times instead of 32 times, all values in the last three columns of the above table would be multiplied by a factor of $\frac{N}{32}$, which would then cancel out in the remaining calculations.

EXAMPLE

Three dice are rolled 216 times. Construct a binomial distribution and show the frequencies of occurrence for 0,1,2 and 3 sixes.

Evaluate the mean and the standard deviation of the distribution.

Solution

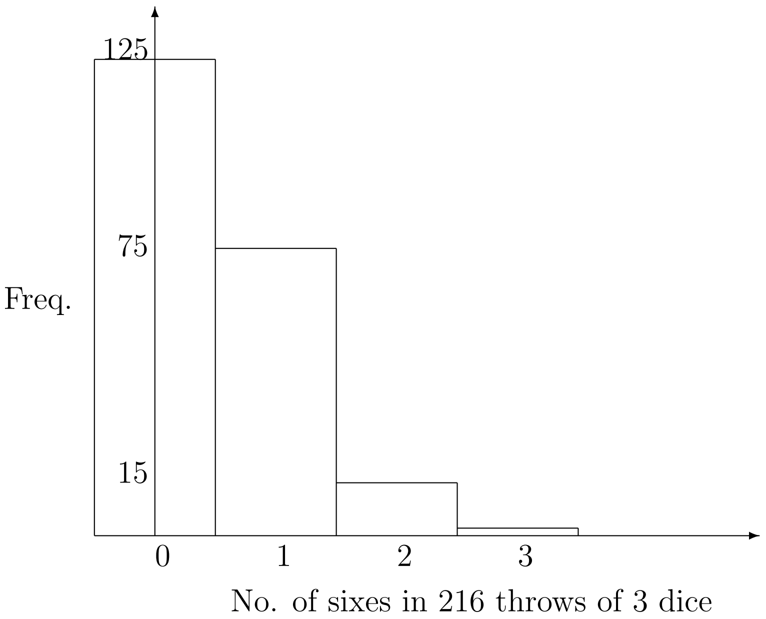
First of all, the probability of success in obtaining a six with a single throw of a die is $\frac{1}{6}$, and the corresponding probability of failure is $\frac{5}{6}$.

For a single throw of three dice, we require the expansion

$$\left(\frac{1}{6}+\frac{5}{6}\right)^3 \equiv \left(\frac{1}{6}\right)^3 + 3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) + 3\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 + \left(\frac{5}{6}\right)^3,$$

showing that the probabilities of 0,1,2 and 3 sixes are $\frac{125}{216}$, $\frac{75}{216}$, $\frac{15}{216}$ and $\frac{1}{216}$, respectively.

Hence, in 216 throws of the three dice we may expect 0 sixes, 125 times; 1 six, 75 times; 2 sixes, 15 times and 3 sixes, once. The corresponding histogram is as follows:



From the previous Theorem, the mean value is

$$3 \times \frac{1}{6} = \frac{1}{2}$$

and the standard deviation is

$$\sqrt{3 \times \frac{1}{6} \times \frac{5}{6}} = \frac{\sqrt{15}}{6}.$$

19.6.3 EXERCISES

1. Four dice are rolled 81 times. If less than 5 on a die is considered to be a success, and everything else a failure,

   (a) draw the corresponding histogram for the expected frequencies of success;

   (b) determine the expected number of times of obtaining at least three successes among the four dice;

   (c) shade the area of the histogram which is a measure of the result in (c);

   (d) calculate the mean and the standard deviation of the frequency distribution in (a).

2. In a seed-viability test, 450 seeds were placed on a filter-paper in 90 rows of 5. The number of seeds that germinated in each row were counted, and the results were as follows:

No. of seeds germinating per row	0	1	2	3	4	5
Observed Frequency of rows	0	1	11	30	38	10

   If the germinating seeds were distributed, at random, among the rows, we would expect a binomial distribution with an index of $n = 5$.

   Determine

   (a) the average number of seeds germinating per row;

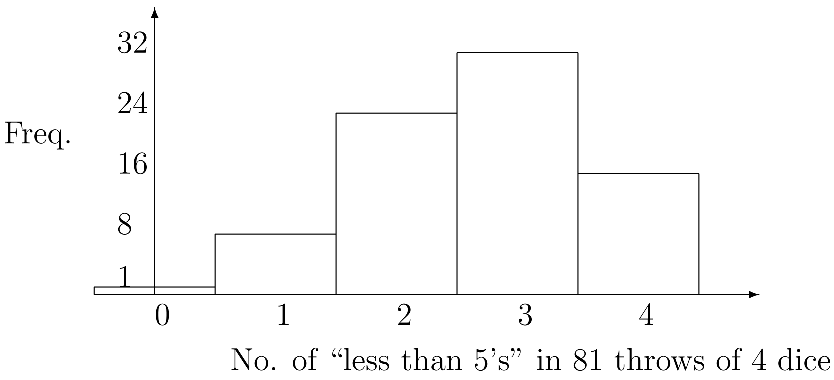
   (b) the probability of a single seed germinating;

   (c) the expected frequencies of rows for each number of seeds germinating;

   Draw the histogram for the expected frequencies and the histogram for the observed frequencies.

19.6.4 ANSWERS TO EXERCISES

1. (a) The histogram is as follows:



   (b) Expected frequency of 3 or 4 successes $= 32 + 16 = 48$;

   (c) Shade the last two rectangles on the right of the histogram;

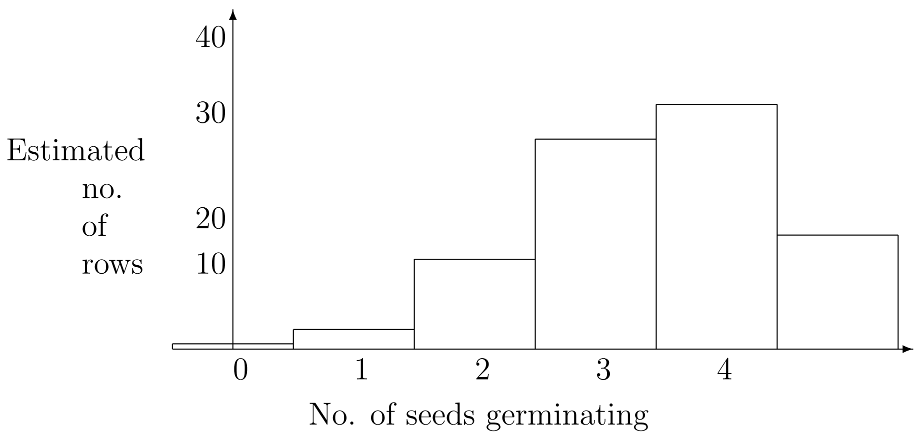
   (d) Mean $= 4 \times \frac{2}{3} = \frac{8}{3}$ and Standard Deviation $= \sqrt{4 \times \frac{2}{3} \times \frac{1}{3}} = \frac{2\sqrt{2}}{3}$.

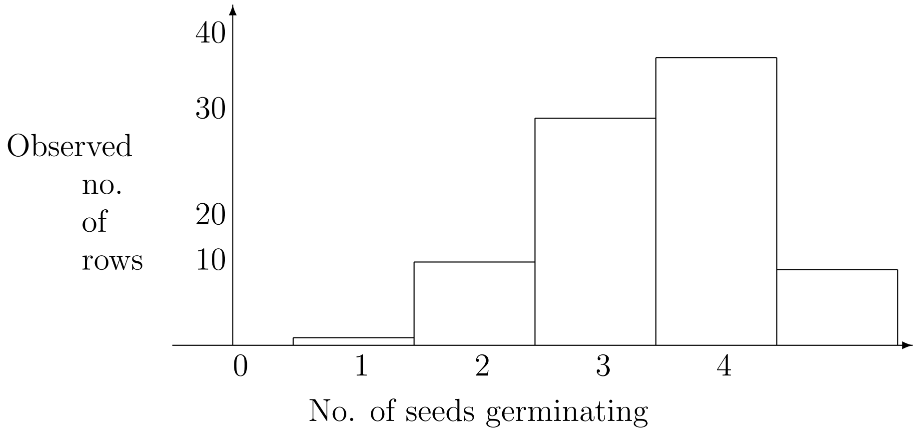
2. (a) Average number of seeds germinating per row is $\frac{325}{90} = 3.50$;

   (b) Probability of a single seed germinating is $\frac{3.5}{5} = \frac{7}{10}$;

   (c) Expected frequencies for 0,1,2,3,4,5 seeds are 0.2187, 2.5515, 11.9070, 27.7830, 32.4135, 15.1263 respectively.

   (d) The histograms are as follows:





Observed no. of rows
40
30
20
10
0
1
2
3
4
No. of seeds germinating