3

Some applications of the hypergeometric and Poisson distributions

In this chapter we will consider some practical situations where the hypergeometric and Poisson distributions arise. We will first consider a technique for estimating animal populations known as **capture–recapture.** This, as we shall see, involves the **hypergeometric distribution.** Poisson random variables arise when we consider randomly distributed points in space or time. One of the applications of this is in the analysis of spatial patterns of plants, which is important in **forestry.** Finally we consider **compound Poisson random variables** with a view to analysing some experimental results in **neurophysiology.**

3.1 THE HYPERGEOMETRIC DISTRIBUTION

The hypergeometric distribution is obtained as follows. A sample of size n is drawn, without replacement, from a population of size N composed of M individuals of type 1 and N–M individuals of type 2. Then the number X of individuals of type 1 in the sample is a **hypergeometric random variable** with probability mass function

$$p_k = \Pr\{X = k\} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad \max(0, n-N+M) \leqslant k \leqslant \min(M, n). \tag{3.1}$$

$n-(N-M)$

To derive (3.1) we note that there are $\binom{N}{n}$ ways of choosing the sample of size n from N individuals. The k individuals of type 1 can be chosen from M in $\binom{M}{k}$ ways, and the $n-k$ individuals of type 2 can be chosen from $N-M$ in $\binom{N-M}{n-k}$ ways. Hence there are $\binom{M}{k}\binom{N-M}{n-k}$ distinct samples with k individuals of type 1 and so (3.1) gives the proportion of samples of size n which contain k individuals of type 1.

The range of X is as indicated in (3.1) as the following arguments show. Recall that there are $N-M$ type 2 individuals. If $n \leqslant N-M$ all members of the sample can be type 2 so it is possible that there are zero type 1 individuals. However, if $n > N-M$, there must be some, and in fact at least $n-(N-M)$, type 1 individuals in the sample. Thus the smallest possible value of X is the larger of 0 and $n-N+M$. Also, there can be no more than n individuals of type 1 if $n \leqslant M$ and no more than M if $M \leqslant n$. Hence the largest possible value of X is the smaller of M and n.

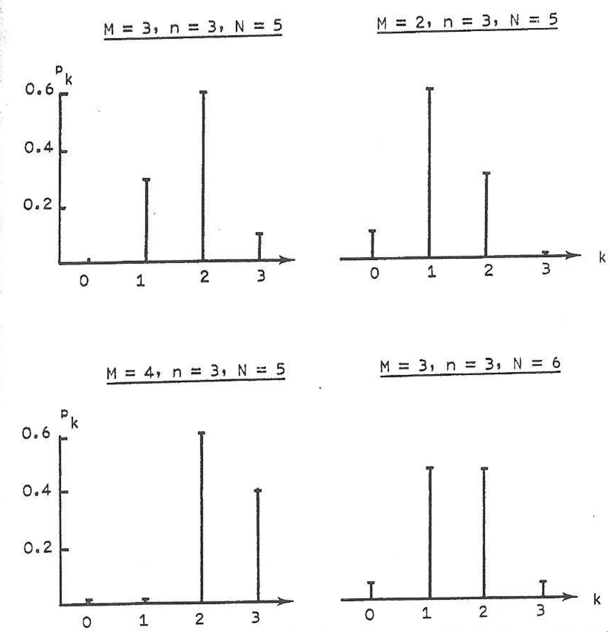

Figure 3.1 Probability mass functions for hypergeometric distributions with various values of the parameters N, M and n.

For the curious we note that the distribution is called hypergeometric because the corresponding generating function is a hypergeometric series (Kendall and Stuart, 1958).

The shape of the hypergeometric distribution depends on the values of the parameters N, M and n. Some examples for small parameter values are shown in Fig. 3.1. Tables are given in Liebermann and Owen (1961).

$G(z) = E(z^X) = \sum p(x) z^x$

Mean and variance

The mean of X is

$$E(X) = \frac{nM}{N}$$

and its variance is

$$\mathrm{Var}(X) = \frac{nM(N-n)(N-M)}{N^2(N-1)}. \quad (3.2)$$

Proof We follow the method of Moran (1968). Introduce the **indicator random variables** defined as follows. Let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th member of the sample is type 1,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the total number of type 1 individuals is

$$X = \sum_{i=1}^{n} X_i.$$

Each member of the sample has the same probability of being type 1. Indeed,

$$\Pr\{X_i = 1\} = \frac{M}{N}, \quad i = 1, \ldots, n. \quad (3.3)$$

as follows from the probability law (3.1) when $k = n = 1$.

Since

$$E(X_i) = \frac{M}{N},$$

we see that

$$E(X) = nE(X_i) = \frac{nM}{N}.$$

To find $\mathrm{Var}(X)$ we note that the second moment of X is

$$E(X^2) = E\left(\left(\sum_{i=1}^{n} X_i\right)^2\right) = E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right)$$

$$= E\left(\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n} X_i X_j\right). \quad (3.4)$$

The expected value of X_i^2 is, from (3.3),

$$E(X_i^2) = \frac{M}{N}, \quad i = 1, 2, \ldots, n. \quad (3.5)$$

We now use (3.1) with $n = k = 2$, to get the probability that $X = 2$ when $n = 2$. This gives

$$\Pr\{X_i = 1, X_j = 1\} = \frac{\binom{M}{2}\binom{N-M}{0}}{\binom{N}{2}}$$

$$= \frac{M(M-1)}{N(N-1)}, \quad i, j = 1, 2, \ldots, n, \quad i \neq j.$$

Thus

$$E(X_i X_j) = \frac{M(M-1)}{N(N-1)}, \quad (3.6)$$

and there are $n(n-1)$ such terms in (3.4). Substituting (3.5) and (3.6) in (3.4) gives

$$E(X^2) = \frac{nM}{N} + \frac{n(n-1)M(M-1)}{N(N-1)}.$$

Formula (3.2) follows from the relation

$$\mathrm{Var}(X) = E(X^2) - E^2(X).$$

3.2 ESTIMATING A POPULATION FROM CAPTURE–RECAPTURE DATA

Assume now that a population size N is unknown. The population may be the kangaroos or emus in a certain area or perhaps the fish in a lake or stream. We wish to estimate N without counting every individual. A method of doing this is called capture–recapture. Here M individuals are captured, marked in some distinguishing way and then released back into the population. Later, after a satisfactory mixing of the marked and unmarked individuals is attained, a sample of size n is taken from the population and the number X of marked individuals is noted. This method, introduced by Laplace in 1786 to estimate France's population, is often employed by biologists and individuals in resource management to estimate animal populations. The method we have described is called **direct sampling**. Another method (**inverse sampling**) is considered in Exercises 5 and 6.

In the capture–recapture model the marked individuals correspond to type 1 and the unmarked to type 2. The probability that the number of marked individuals is k is thus given by (3.1). Suppose now that k is the number of marked individuals in the sample; then values of N for which $\Pr\{X=k\}$ is very small are considered unlikely. One takes as an estimate, $\hat{N}$, of N that value which maximizes $\Pr\{X=k\}$. $\hat{N}$ is a random variable and is called the **maximum likelihood estimate** of N.

Theorem 3.1 The maximum likelihood estimate of N is

$$\hat{N}=\left[\frac{Mn}{X}\right],$$

where $[z]$ denotes the greatest integer less than z.

Proof We follow Feller (1968). Holding M and n constant we let $\Pr\{X=k\}$ for a fixed value of N be denoted $p_N(k)$. Then

$$\frac{p_N(k)}{p_{N-1}(k)}=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\Bigg/\frac{\binom{M}{k}\binom{N-1-M}{n-k}}{\binom{N-1}{n}},$$

which simplifies to

$$\frac{p_N(k)}{p_{N-1}(k)}=\frac{N^2-MN-nN+Mn}{N^2-MN-nN+kN}.$$

We see that p_N is greater than, equal to, or less than p_{N-1} according as Mn is greater than, equal to, or less than kN; or equivalently as N is less than, equal to or greater than Mn/k. Excluding for now the case where Mn/k is an integer, the sequence $\{p_N, N=1,2,\ldots\}$ is increasing as long as $N<Mn/k$ and is decreasing when $N>Mn/k$. Thus the maximum value of p_N occurs when

$$N=\left[\frac{Mn}{k}\right],$$

which is the largest integer less than Mn/k.

In the event that Mn/k is an integer the maximum value of p_N will be $p_{Mn/k}$ and $p_{Mn/k-1}$, these being equal. One may then use

$$\frac{Mn}{k}-1=\left[\frac{Mn}{k}\right]$$

as an estimate of the population. This completes the proof.

Approximate confidence intervals for $\hat{N}$

In situations of practical interest N will be much larger than both M and n. Let us assume in fact that N is large enough to regard the sampling as approximately with replacement. If $\tilde{X}_i$ approximates X_i in this scheme, then, for all i from 1 to n,

$$\Pr\{\tilde{X}_i=1\}=\frac{M}{N}=1-\Pr\{\tilde{X}_i=0\}.$$

The approximation to X is then given by

$$\tilde{X}=\sum_{i=1}^{n}\tilde{X}_i.$$

This is a binomial random variable with parameters n and M/N so that

$$\Pr\{\tilde{X}=k\}=b\left(k;n,\frac{M}{N}\right),\qquad k=0,1,\ldots,n.$$

The expectation and variance of $\tilde{X}$ are

$$E(\tilde{X})=\frac{nM}{N}=E(X)$$

$$\operatorname{Var}(\tilde{X})=\frac{nM}{N}\left(1-\frac{M}{N}\right)\simeq\operatorname{Var}(X).$$

Furthermore, if the sample size n is fairly large, the distribution of $\tilde{X}$ can be approximated by that of a normal random variable with the same mean and variance (see Chapter 6). Replacing N by the observed value, $\hat{n}$, of its maximum likelihood estimator gives

$$\tilde{X}\overset{d}{\simeq}N\left(\frac{nM}{\hat{n}},\sqrt{\frac{nM}{\hat{n}}\left(1-\frac{M}{\hat{n}}\right)}\right),$$

where $\overset{d}{\simeq}$ means 'has approximately the same distribution'.

Ignoring the technicality of integer values, we have

$$\hat{n}=\frac{nM}{k},$$

where k is the observed value of X, so

$$\tilde{X} \overset{d}{\simeq} N\left(k, \sqrt{k\left(1-\frac{k}{n}\right)}\right).$$

Using the standard symbol Z for an $N(0, 1)$ random variable and the usual notation

$$\Pr\{Z > z_{\alpha/2}\} = \frac{\alpha}{2},$$

we find

$$\Pr\left\{k - z_{\alpha/2}\sqrt{k\left(1-\frac{k}{n}\right)} < \tilde{X} < k + z_{\alpha/2}\sqrt{k\left(1-\frac{k}{n}\right)}\right\} \simeq 1-\alpha.$$

However, $\hat{N} = Mn/X$, so we obtain the following approximate result when the observed number of marked individuals in the recaptured sample is k.

Theorem 3.2 An approximate $100(1-\alpha)\%$ confidence interval for the estimator $\hat{N}$ of the population is

$$\Pr\left\{\frac{nM}{k + z_{\alpha/2}\sqrt{k\left(1-\frac{k}{n}\right)}} < \hat{N} < \frac{nM}{k - z_{\alpha/2}\sqrt{k\left(1-\frac{k}{n}\right)}}\right\} \simeq 1-\alpha.$$

Thus for example, if a 95% confidence interval is required we put $z_{\alpha/2} = z_{.025} = 1.96$ in this formula.

Discussion

The above estimates have been obtained for direct sampling in the ideal situation. Before applying them in any real situation an examination of the assumptions made would be worth while. Among these are:

(i) The marked individuals disperse randomly and homogeneously throughout the population.
(ii) All marked individuals retain their marks.
(iii) Each individual, whether marked or not, has the same chance of being in the recaptured sample.
(iv) There are no losses due to death or emigration and no gains due to birth or immigration.

Some of these assumptions can be relaxed in a relatively simple way (see Exercise 7). In the approach mentioned earlier called inverse sampling, the recapturing takes place until a predetermined number of marked individuals is obtained. For useful refinements of the basic method presented above see Manly (1984) and references therein; see also Cormack (1968) and the conference article of the same author (1973) who begins with the following remarks:

> Many of the papers in this volume are concerned with the process of describing the development of an animal population by a mathematical model. The properties of such a model can then be derived, either by elegant mathematics or equally elegant computer simulation, in order to describe the future state of the population in terms of certain initial boundary conditions. The model becomes of scientific value when such predictions can be tested, which requires in turn that the mathematical symbols can be replaced by numbers. The parameters of the model must be estimated from data of a type that a biologist can collect about the population he is studying.

For an introductory treatment written for biologists, see Begon (1979).

3.3 THE POISSON DISTRIBUTION

We recall the definition and some elementary properties of Poisson random variables.

Definition A non-negative integer-valued random variable X has a Poisson distribution with parameter $\lambda > 0$ if

$$p_k = \Pr\{X = k\} = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots \tag{3.7}$$

From the definition of e^{λ} as $\sum_0^{\infty} \lambda^k/k!$ we find

$$\sum_{k=0}^{\infty} \Pr\{X = k\} = 1.$$

The mean and variance of X will easily be found to be

$$E(X) = \mathrm{Var}(X) = \lambda.$$

The shape of the probability mass function depends on λ as Table 3.1 and the graphs of Fig. 3.2 illustrate.

Table 3.1 Probability mass functions for some Poisson random variables

	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
$\lambda = \frac{1}{2}$	.607	.303	.076	.013	.002	<.001			
$\lambda = 1$	.368	.368	.184	.061	.015	.003	<.001		
$\lambda = 2$	.135	.271	.271	.180	.090	.036	.012	.003	<.001