Hypergeometric Distribution and Its Application in Statistics

Chapter · January 2011

DOI: 10.1007/978-3-642-04898-2_294

1 author:

Anwar Joarder
University of Liberal Arts Bangladesh (ULAB)
119 PUBLICATIONS **558** CITATIONS

Hypergeometric Distribution and Its Application in Statistics

An important discrete distribution encountered in sampling situations is the hypergeometric distribution. Suppose that a finite population of Nitemscontainstwotypesof itemsinwhichKitemsareof onekind(saydefective)andN − K items are of a different kind (say non-defective). If n items are drawn at random in succession, without replacement, then X denoting the number of defective items selected follows a hypergeometric distribution. The probability of the event $D_1 D_2 \cdots D_x D'_{x+1} \cdots D'_n$ denoting x successive defectives items and n − x successive non-defective items is given by

$$P(D_1 D_2 \cdots D_x D'_{x+1} \cdots D'_n) = \frac{C_{K-x}^{N-n}}{C_K^N}, \quad \max\{0, n-(N-K)\} \le x \le \min\{n, K\}, \tag{1}$$

where C_x^n is the number of combinations of x items that can be chosen from a group of n items and is equal to n! / [x!(n − x)!]. The probability of any other particular sequence in the sample space is also the same as (1). Interested readers may refer to Joarder and Al-Sabah (2007).

Since there are C_x^n outcomes having x defective items and (n − x) non-defective items out of at most 2^n elements in the sample space, the probability of x successes in n trials is given by

$$P(X = x) = \frac{C_x^n C_{K-x}^{N-n}}{C_K^N}, \quad \max\{0, n-(N-K)\} \le x \le \min\{n, K\}, \tag{2}$$

(cf. Kendall and Stuart 1969, p. 133). The probability of x successes in n trials is commonly written as

$$P(X = x) = \frac{C_x^K C_{n-x}^{N-K}}{C_n^N}, \quad \max\{0, n-(N-K)\} \le x \le \min\{n, K\}. \tag{3}$$

Vandermonde's identity justifies the equivalence of the two forms in (2) and (3). The proof of (3) is available in most textbooks on statistics (e.g., Johnson 2007) and discrete mathematics (e.g., Barnett 1998). There are C_x^K ways of choosing x of the K items (say defective items) and C_{n-x}^{N-K} ways of choosing (n − x) of the (N − K) non-defective items, and hence there are $C_x^K C_{n-x}^{N-K}$ ways of choosing x defectives and (n − x) non-defective items. Since there are C_n^N ways of choosing n of the N elements, assuming C_n^N sample points are equally likely, the probability of having x defective items in the sample is given by (3).

The name hypergeometric is derived from a series introduced by the Swiss mathematician and physicist, Leonard Euler, in 1769. The probabilities in (3) are the successive terms of

$$\frac{(N-n)!(N-K)!}{N!(N-K-n)!}\ {}_2F_1(-n,-K;N-K-n+1;1), \tag{4}$$

where ${}_2F_1(a_1, a_2; b; x)$ is the generalized hypergeometric function (Johnson et al. 1993, p. 237).

Suppose that a random sample of n = 3 items is selected from a lot of N = 5 items in which there are K = 3 defective items (distinguishable or indistinguishable) and 2 non-defective items (distinguishable or indistinguishable). Let D_i (i = 1, 2, 3) be the event that we have a defective item in the ith selection, and N_i (i = 1, 2, 3) be the event that we have a non-defective item in the ith selection. Also let X be the number of defective items selected in the sample. The elements of the sample space are given by $D_1 D_2 D_3$, $D_1 D_2 N_3$, $D_1 N_2 D_3$, $D_1 N_2 N_3$, $N_1 D_2 D_3$, $N_1 D_2 N_3$, and $N_1 N_2 D_3$. By (1), the probabilities are given by 0.10, 0.2, 0.2, 0.10, 0.2, 0.10 and 0.10 respectively. Note that elements in the

sample space are not equiprobable. The probability of having two defective items in the sample is given by P(X = 2) = 0.2 + 0.2 + 0.2 = 0.6. If n is large, it is not feasible to write out the sample space but one can use (2) directly.

Note that if the items in each of the two categories are distinguishable, or labeled to make them distinguishable, the sample space can be written out with all distinguishable items. Then the sample outcomes are equally likely or equiprobable resulting in a Simple Random Sampling. In the above example, let the defective items be labeled as D^1, D^2 and D^3 while the non-defective items be labeled as N^1 and N^2 to make the items in the population distinguishable. Then $C_n^N = 10$ elements in the sample space are given by $D^1 D^2 D^3$, $D^1 D^2 N^1$, $D^1 D^2 N^2$, $D^1 D^3 N^1$, $D^1 D^3 N^2$, $D^1 N^1 N^2$, $D^2 D^3 N^1$, $D^2 D^3 N^2$, $D^2 N^1 N^2$, and $D^3 N^1 N^2$ each with the probability 0.10, and hence P(X = 2) = 0.6. In case C_n^N is large, it is not feasible to write out the sample space but one can use (3) directly.

Suppose that n items are drawn at random, with replacement, and X denotes the number of defective items selected. The probability that any item is defective at any draw is p = K / N (say). Then with arguments similar to above, the probability of having x defectives and (n − x) non-defectives in any of the C_x^n sequence is given by $p^x q^{n-x}$ so that $P(X = x) = C_x^n p^x q^{n-x}$. Now if $N \to \infty$, and p = K / N, it is easy to prove that (2) has a limiting value of $C_x^n p^x q^{n-x}$. This shows the equivalence of binomial and hypergeometric distribution in the limit.

The mean and variance of hypergeometric distribution are given by np and (1 − f)npq respectively, where p = K / N, q = 1 − p, and f is the finite population correction factor defined by (N − 1)f = N − n. The mode of the distribution is the greatest integer not exceeding $\frac{(n+1)(K+1)}{N+2}$. The coefficient of skewness and that of kurtosis are given by

$$\frac{(N-2K)(N-2n)(N-1)^{1/2}}{[nK(N-K)(N-n)]^{1/2}(N-2)}, \tag{5}$$

and

$$\frac{N^2(N-1)}{n(N-2)(N-3)(N-n)}\left[\frac{N(N+1)-6n(N-n)}{K(N-K)} + \frac{3n(N-n)(N+6)}{N^2} - 6\right], \tag{6}$$

respectively (Evans et al. 2000, p. 111).

The maximum likelihood estimator of the number of defectives K in a lot is the greatest integer not exceeding x(N + 1) / n; if x(N + 1) / n is an integer, then [x(N + 1) / n] − 1 also maximizes the likelihood function (Johnson et al. 1993, p. 263).

The distribution has got a number of important applications in the real world. In the industrial quality control, lots of size N containing a proportion of p defectives are sampled by using samples of fixed size n. The number of defectives X per sample follows a hypergeometric distribution. If X ≤ c (the acceptance number), the lot is accepted; otherwise it is rejected. The design of suitable sampling plans requires the calculation of confidence intervals of Np, given N, n and c. Tables of these have been published by Chung and DeLury (1950) and Owen (1962). It is worth mentioning that in many cases binomial or Poisson approximations to the hypergeometric distribution suffice.

Another useful application is the estimation of the size of the animal populations from capture-recapture data. This kind of application dates back to Peterson (1896), quoted by Chapman (1952). Consider, for example, the estimation of the number N of animals in a population. A sample of size K is captured, tagged and then released into the population. After a while a new catch of n animals is made, the number of tagged animals (X) in the sample is noted. Assume that the two catches are random samples from the population of all animals. Indeed, if we assume that there were no births or deaths, then the proportion of tagged animals in the sample (X／n) is approximately the same as that in the population (K／N). That is, an estimate of N is \hat{N} = nK／X. It may be noted that this estimate maximizes the probability of observed value of X. Evidently, X has a hypergeometric distribution with probability mass function given by (2) or (3).

The hypergeometric distribution can be approximated by Poisson distribution with parameter λ if K, N and n all tend to

infinity for K/N small and nK/N tending to λ. It can also be approximated by normal distribution if n is large, but x/N is not too small. A concise description of many other types of hypergeometric distribution and their properties are available in Johnson et al. (1993).

About the Author

Dr. Anwar H Joarder has been working at the Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia since 1997. He also worked at the University of Western Ontario, Jahangir Nagar University, University of Dhaka, North South University, Monash University and the University of Sydney. He is an Elected member of the Royal Statistical Society and the International Statistical Institute. He authored and co-authored around 60 research papers. He serves on the editorial board of five journals including Model Assisted Sampling and Applications and International Journal of Mathematical Education in Science and Technology.

Cross References

Exact Inference for Categorical Data
Fisher Exact Test
Minimum Variance Unbiased
Most Powerful Test
Multivariate Statistical Distributions
Poisson Distribution and Its Application in Statistics
Proportions, Inferences, and Comparisons
Statistical Distributions: An Overview
Univariate Discrete Distributions: An Overview

References and Further Reading

- Barnett S (1998) Discrete mathematics: numbers and beyond. Pearson Education Limited, Essex, England
- Chapman DG (1952) Inverse, multiple and sequential sample censuses. Biometrics 8:286-306
- Chung JH, DeLury DB (1950) Confidence limits for the hypergeometric distribution. University of Toronto Press, Toronto, Canada
- Evans M, Hastings N, Pecock B (2000) Statistical distributions. Wiley, New York
- Joarder AH, Al-Sabah WS (2007) Probability issues in without replacement sampling. Int J Math Educ Sci Technol 38(6):823-831
- Johnson R (2007) Miller and Freund's Probability and statistics for engineers. Pearson Educational International, New Jersey, USA
- Johnson NL, Kotz S, Kemp AW (1993) Univariate discrete distributions. Wiley, New York, USA
- Kendall MG, Stuart A (1969) The advanced theory of statistics, vol 1: distribution theory. Charles Griffin, London
- Owen DB (1962) Handbook of statistical tables. Addison-Wesley, Reading, MA
- Peterson CGJ (1896) The yearly immigration of young plaice into the Limfjord from the German sea. Danish Biol Station Rep 6:5-48

Hypergeometric Distribution and Its Application in Statistics

Anwar H. Joarder King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

DOI: 10.1007/SpringerReference_205377
URL: http://www.springerreference.com/index/chapterdbid/205377
Part of: International Encyclopedia of Statistical Science
Editor: Prof. Miodrag Lovric
PDF created on: February, 03, 2012 17:54