3. The Multivariate Hypergeometric Distribution

Basic Theory

As in the basic sampling model, we start with a finite population D consisting of m objects. In this section, we suppose in addition that each object is one of k types; that is, we have a **multi-type population**. For example, we could have an urn with balls of several different colors, or a population of voters who are either *democrat*, *republican*, or *independent*. Let D_i denote the subset of all type i objects and let $m_i = \#(D_i)$ for $i \in \{1, 2, \ldots, k\}$. Thus

$$D = \bigcup_{i=1}^k D_i, \quad m = \sum_{i=1}^k m_i$$

The dichotomous model considered earlier is clearly a special case, with $k = 2$. As in the basic sampling model, we sample n objects at random from D:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$$

where $X_i \in D$ is the i^{th} object chosen. Now let Y_i denote the number of type i objects in the sample, for $i \in \{1, 2, \ldots, k\}$. Note that

$$\sum_{i=1}^k Y_i = n$$

so if we know the values of $k - 1$ of the counting variables, we can find the value of the remaining counting variable. As with any counting variable, we can express Y_i as a sum of indicator variables:

1. Show that

$$Y_i = \sum_{j=1}^n I_{i,j} \quad \text{where } I_{i,j} = \begin{cases} 1, & X_j \in D_i \\ 0, & X_j \notin D_i \end{cases}$$

We assume initially that the sampling is without replacement, since this is the realistic case in most applications.

The Joint Distribution

Basic combinatorial arguments can be used to derive the probability density function of the random vector of counting variables. Recall that since the sampling is without replacement, the unordered sample is uniformly distributed over the combinations of size n chosen from D.

2. Show that

$$\mathbb{P}(Y_1 = j_1, Y_2 = j_2, \ldots, Y_k = j_k) = \frac{\binom{m_1}{j_1}\binom{m_2}{j_2} \cdots \binom{m_k}{j_k}}{\binom{m}{n}} \quad \text{for } (j_1, j_2, \ldots, j_k) \in \mathbb{N}^k \text{ with } \sum_{i=1}^k j_i = n$$

The distribution of $(Y_1, Y_2, \ldots, Y_k)$ is called the **multivariate hypergeometric distribution** with parameters m, $(m_1, m_2, \ldots, m_k)$, and n. We also say that $(Y_1, Y_2, \ldots, Y_{k-1})$ has this distribution (recall again that the values of any $k - 1$ of the variables determines the value of the remaining variable). Usually it is clear from context which meaning is intended. The ordinary hypergeometric distribution corresponds to $k = 2$.

3. Show the following alternate from of the multivariate hypergeometric probability density function in two ways: *combinatorially*, by considering the ordered sample uniformly distributed over the permutations of size n chosen from D, and *algebraically*, starting with the result in Exercise 2.

$$\mathbb{P}(Y_1 = j_1, Y_2 = j_2, \ldots, Y_k = j_k) = \binom{n}{j_1, j_2, \ldots, j_k} \frac{m_1^{(j_1)} m_2^{(j_2)} \cdots m_k^{(j_k)}}{m^{(n)}} \quad \text{for } (j_1, j_2, \ldots, j_k) \in \mathbb{N}^k \text{ with } \sum_{i=1}^k j_i = n$$

The Marginal Distributions

4. Show that Y_i has the hypergeometric distribution with parameters m, m_i. and n. Give both a probabilistic argument, based on the sampling model, and an analytic derivation, based on the joint probability density function in Exercise 2.

$$\mathbb{P}(Y_i = j) = \frac{\binom{m_i}{j}\binom{m - m_i}{n - j}}{\binom{m}{n}}, \quad j \in \{0, 1, \ldots, n\}$$

Grouping

The multivariate hypergeometric distribution is preserved when the counting variables are combined. Specifically, suppose that $(A_1, A_2, \ldots, A_l)$ is a partition of the index set $\{1, 2, \ldots, k\}$ into nonempty, disjoint subsets. Let

$$W_j = \sum_{i \in A_j} Y_i, \quad r_j = \sum_{i \in A_j} m_i \quad \text{for } j \in \{1, 2, \ldots, l\}$$

5. Show that $(W_1, W_2, \ldots, W_l)$ has the multivariate hypergeometric distribution with parameters m, $(r_1, r_2, \ldots, r_l)$, and n.

Conditioning

The multivariate hypergeometric distribution is also preserved when some of the counting variables are observed. Specifically, suppose that (A, B) is a partition of the index set $\{1, 2, \ldots, k\}$ into nonempty, disjoint subsets . Suppose that we observe $Y_j = y_j$ for $j \in B$. Let

$$z = \sum_{j \in B} y_j, \quad r = \sum_{i \in A} m_i$$

6. Show that the conditional distribution of $[Y_i : i \in A]$ given $\{Y_j = y_j : j \in B\}$ is multivariate hypergeometric with parameters r, $[m_i : i \in A]$, and z.

Combinations of the basic results in Exercise 5 and Exercise 6 can be used to compute any marginal or conditional distributions of the counting variables.

Moments

We will compute the mean, variance, covariance, and correlation of the counting variables. Results from the hypergeometric distribution and the representation in terms of indicator variables in Exercise 1 are the main tools.

7. Show that for $i \in \{1, 2, ..., k\}$,

a. $\mathbb{E}(Y_i) = n \frac{m_i}{m}$

b. $\text{var}(Y_i) = n \frac{m_i}{m} \left(1 - \frac{m_i}{m}\right) \frac{m-n}{m-1}$

8. Suppose that i and j are distinct elements of $\{1, 2, ..., k\}$ and that r and s are distinct elements of $\{1, 2, ..., n\}$. Show that

a. $\text{cov}\left(I_{i,r}, I_{j,r}\right) = -\frac{m_i m_j}{m^2}$

b. $\text{cov}\left(I_{i,r}, I_{j,s}\right) = -\frac{m_i m_j}{m^2 (m-1)}$

9. Suppose that i and j are distinct elements of $\{1, 2, ..., k\}$ and that r and s are distinct elements of $\{1, 2, ..., n\}$. Show that

a. $\text{cor}\left(I_{i,r}, I_{j,r}\right) = -\frac{m_i m_j}{\sqrt{(m-m_i)(m-m_j)}}$

b. $\text{cor}\left(I_{i,r}, I_{j,s}\right) = -\frac{m_i m_j}{\sqrt{(m-m_i)(m-m_j)}(m-1)}$

In particular, $I_{i,r}$ and $I_{j,s}$ are negatively correlated for distinct i and j, and for any r and s. Does this result seem reasonable?

10. Use the result of Exercise 7 and Exercise 8 to show that for distinct i and j in $\{1, 2, ..., k\}$,

a. $\text{cov}\left(Y_i, Y_j\right) = -n \frac{m_i m_j}{m^2} \frac{m-n}{m-1}$

b. $\cor(Y_i, Y_j) = -\frac{m_i m_j}{\sqrt{(m - m_i)(m - m_j)}}$

Sampling with Replacement

Suppose now that the sampling is with replacement, even though this is usually not realistic in applications.

11. Show that the types of the objects in the sample form a sequence of n multinomial trials with parameters $\left(\frac{m_1}{m}, \frac{m_2}{m}, \ldots, \frac{m_k}{m}\right)$.

The following results now follow immediately from the general theory of multinomial trials, although modifications of the arguments above could also be used.

12. Show that $(Y_1, Y_2, \ldots, Y_k)$ has the multinomial distribution with parameters n and $\left(\frac{m_1}{m}, \frac{m_2}{m}, \ldots, \frac{m_k}{m}\right)$:

$$\P(Y_1 = j_1, Y_2 = j_2, \ldots, Y_k = j_k) = \binom{n}{j_1, j_2, \ldots, j_k} \frac{m_1^{j_1} m_2^{j_2} \cdots m_k^{j_k}}{m^n} \quad \text{for } (j_1, j_2, \ldots, j_k) \in \N^k \text{ with } \sum_{i=1}^k j_i = n$$

13. Show that for distinct i and j in $\{1, 2, \ldots, k\}$,

a. $\E(Y_i) = n \frac{m_i}{m}$

b. $\var(Y_i) = n \frac{m_i}{m}\left(1 - \frac{m_i}{m}\right)$

c. $\cov(Y_i, Y_j) = -n \frac{m_i m_j}{m^2}$

d. $\cor(Y_i, Y_j) = -\frac{m_i m_j}{\sqrt{(m - m_i)(m - m_j)}}$

Convergence to the Multinomial Distribution

Suppose that the population size m is very large compared to the sample size n. In this case, it seems reasonable that sampling *without* replacement is not too much different than sampling *with* replacement, and hence the multivariate hypergeometric distribution should be well approximated by the multinomial. The following exercise makes this observation precise. Practically, it is a valuable result, since in many cases we do not know the population size exactly.

14. Suppose that m_i depends on m and that $\frac{m_i}{m} \to p_i$ as $m \to \infty$ for $i \in \{1, 2, \ldots, k\}$. Show that for fixed n, the multivariate hypergeometric probability density function with parameters m, $(m_1, m_2, \ldots, m_k)$, and n converges to the multinomial probability density function with parameters n and $(p_1, p_2, \ldots, p_k)$. *Hint*: Use the representation in Exercise 3.

Examples and Applications

15. A population of 100 voters consists of 40 republicans, 35 democrats and 25 independents. A random sample of 10 voters is chosen.

a. Find the joint density function of the number of republicans, number of democrats, and number of independents in the sample
b. Find the mean of each variable in (a).
c. Find the variance of each variable in (a).
d. Find the covariance of each pair of variables in (a).
e. Find the probability that the sample contains at least 4 republicans, at least 3 democrats, and at least 2 independents.

Cards

Recall that the general **card experiment** is to select n cards at random and without replacement from a standard deck of 52 cards. The special case $n = 5$ is the **poker experiment** and the special case $n = 13$ is the **bridge experiment**.

16. In a bridge hand, find the probability density function of

a. the number of spades, number of hearts, and number of diamonds.
b. the number of spades and number of hearts.
c. the number of spades.
d. the number of red cards and the number of black cards.

17. In a bridge hand,

a. Find the mean and variance of the number of spades.
b. Find the covariance and correlation between the number of spades and the number of hearts.
c. Find the mean and variance of the number of red cards.

18. In a bridge hand,

a. Find the conditional probability density function of the number of spades and the number of hearts, given that the hand has 4 diamonds.
b. Find the conditional probability density function of the number of spades given that the hand has 3 hearts and 2 diamonds.

In the card experiment, a hand that does not contain any cards of a particular suit is said to be **void** in that suit.

19. Use the inclusion-exclusion rule to show that the probability that a poker hand is void in at least one suit is

$$\frac{1913496}{2598960} \approx 0.736$$

20. In the card experiment, set $n = 5$. Run the simulation 1000 times, updating after each run. Compute the relative frequency of the event that the hand is void in at least one suit. Compare the relative frequency with the true probability given in the previous exercise.

21. Use the inclusion-exclusion rule to show that the probability that a bridge hand is void in at least one suit is

$$\frac{32427298180}{635013559600} \approx 0.051$$

Virtual Laboratories > 12. Finite Sampling Models > 1 2 3 4 5 6 7 8 9
Contents | Applets | Data Sets | Biographies | External Resources | Keywords | Feedback | ©