12.3: The Multivariate Hypergeometric Distribution

Basic Theory

The Multitype Model

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 *multitype 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 $D = \bigcup_{i=1}^k D_i$ and $m = \sum_{i=1}^k m_i$. 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. Thus the outcome of the experiment is $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ where $X_i \in D$ is the ith 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:

For $i \in \{1, 2, \ldots, k\}$

$$Y_i = \sum_{j=1}^n \mathbf{1}(X_j \in D_i) \tag{12.3.1}$$

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.

The probability density funtion of $(Y_1, Y_2, \ldots, Y_k)$ is given by

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \frac{\binom{m_1}{y_1}\binom{m_2}{y_2} \cdots \binom{m_k}{y_k}}{\binom{m}{n}}, \quad (y_1, y_2, \ldots, y_k) \in \mathbb{N}^k \text{ with } \sum_{i=1}^k y_i = n \tag{12.3.2}$$

Proof

The binomial coefficient $\binom{m_i}{y_i}$ is the number of unordered subsets of D_i (the type i objects) of size y_i. The binomial coefficient $\binom{m}{n}$ is the number of unordered samples of size n chosen from D. Thus the result follows from the multiplication principle of combinatorics and the uniform distribution of the unordered sample

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 $k = 2$.

An alternate form of the probability density function of $Y_1, Y_2, \ldots, Y_k)$ is

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \binom{n}{y_1, y_2, \ldots, y_k} \frac{m_1^{(y_1)} m_2^{(y_2)} \cdots m_k^{(y_k)}}{m^{(n)}}, \quad (y_1, y_2, \ldots, y_k) \in \mathbb{N}_k \text{ with } \sum_{i=1}^k y_i = n \tag{12.3.3}$$

Combinatorial Proof

The combinatorial proof is to consider the ordered sample, which is uniformly distributed on the set of permutations of size n from D. The multinomial coefficient on the right is the number of ways to partition the index set $\{1, 2, \ldots, n\}$ into k groups where group i has y_i elements (these are the coordinates of the type i objects). The number of (ordered) ways to select the type i objects is $m_i^{(y_i)}$. The denominator $m^{(n)}$ is the number of ordered samples of size n chosen from D.

Algebraic Proof

There is also a simple algebraic proof, starting from the first version of probability density function above. Write each binomial coefficient $\binom{a}{j} = a^{(j)} / j!$ and rearrange a bit.

The Marginal Distributions

For $i \in \{1, 2, \ldots, k\}$, Y_i has the hypergeometric distribution with parameters m, m_i, and n

$$\mathbb{P}(Y_i = y) = \frac{\binom{m_i}{y}\binom{m-m_i}{n-y}}{\binom{m}{n}}, \quad y \in \{0, 1, \ldots, n\} \tag{12.3.4}$$

Proof

An analytic proof is possible, by starting with the first version or the second version of the joint PDF and summing over the unwanted variables. However, a probabilistic proof is much better: Y_i is the number of type i objects in a sample of size n chosen at random (and without replacement) from a population of m objects, with m_i of type i and the remaining $m - m_i$ not of this type.

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

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

Proof

Again, an analytic proof is possible, but a probabilistic proof is much better. Effectively, we now have a population of m objects with l types, and r_i is the number of objects of the new type i. As before we sample n objects without replacement, and W_i is the number of objects in the sample of the new type i.

Note that the marginal distribution of Y_i given above is a special case of grouping. We have two types: type i and not type i. More generally, the marginal distribution of any subsequence of $(Y_1, Y_2, \ldots, Y_n)$ is hypergeometric, with the appropriate parameters.

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 $z = n - \sum_{j \in B} y_j$ and $r = \sum_{i \in A} m_i$.

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.

Proof

Once again, an analytic argument is possible using the definition of conditional probability and the appropriate joint distributions. A probabilistic argument is much better. Effectively, we are selecting a sample of size z from a population of size r, with m_i objects of type i for each $i \in A$.

Combinations of the grouping result and the conditioning result 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 are the main tools.

For $i \in \{1, 2, \ldots, k\}$,

1. $\mathbb{E}(Y_i) = n\frac{m_i}{m}$
2. $\operatorname{var}(Y_i) = n\frac{m_i}{m}\frac{m-m_i}{m}\frac{m-n}{m-1}$

Proof

This follows immediately, since Y_i has the hypergeometric distribution with parameters m, m_i, and n.

Now let $I_{ti} = \mathbf{1}(X_t \in D_i)$, the indicator variable of the event that the tth object selected is type i, for $t \in \{1, 2, \ldots, n\}$ and $i \in \{1, 2, \ldots, k\}$.

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

$$\operatorname{cov}(I_{ri}, I_{rj}) = -\frac{m_i}{m}\frac{m_j}{m} \tag{12.3.5}$$

$$\operatorname{cov}(I_{ri}, I_{sj}) = \frac{1}{m-1}\frac{m_i}{m}\frac{m_j}{m} \tag{12.3.6}$$

Proof

Recall that if A and B are events, then $\text{cov}(A, B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$. In the first case the events are that sample item r is type i and that sample item r is type j. These events are disjoint, and the individual probabilities are $\frac{m_i}{m}$ and $\frac{m_j}{m}$. In the second case, the events are that sample item r is type i and that sample item s is type j. The probability that both events occur is $\frac{m_i}{m} \frac{m_j}{m-1}$ while the individual probabilities are the same as in the first case.

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

$$\text{cor}(I_{ri}, I_{rj}) = -\sqrt{\frac{m_i}{m - m_i} \frac{m_j}{m - m_j}} \tag{12.3.7}$$

$$\text{cor}(I_{ri}, I_{sj}) = \frac{1}{m-1} \sqrt{\frac{m_i}{m - m_i} \frac{m_j}{m - m_j}} \tag{12.3.8}$$

Proof

This follows from the previous result and the definition of correlation. Recall that if I is an indicator variable with parameter p then $\text{var}(I) = p(1 - p)$.

In particular, I_{ri} and I_{rj} are negatively correlated while I_{ri} and I_{sj} are positively correlated.

For distinct i, $j \in \{1, 2, \ldots, k\}$,

$$\text{cov}(Y_i, Y_j) = -n \frac{m_i}{m} \frac{m_j}{m} \frac{m - n}{m - 1} \tag{12.3.9}$$

$$\text{cor}(Y_i, Y_j) = -\sqrt{\frac{m_i}{m - m_i} \frac{m_j}{m - m_j}} \tag{12.3.10}$$

Sampling with Replacement

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

The types of the objects in the sample form a sequence of n multinomial trials with parameters $(m_1/m, m_2/m, \ldots, m_k/m)$

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

$(Y_1, Y_2, \ldots, Y_k)$ has the multinomial distribution with parameters n and $(m_1/m, m_2, /m, \ldots, m_k/m)$

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k) = \binom{n}{y_1, y_2, \ldots, y_k} \frac{m_1^{y_1} m_2^{y_2} \cdots m_k^{y_k}}{m^n}, \quad (y_1, y_2, \ldots, y_k) \in \mathbb{N}^k \text{ with } \sum_{i=1}^k y_i = n \tag{12.3.11}$$

For distinct i, $j \in \{1, 2, \ldots, k\}$,

1. $\mathbb{E}(Y_i) = n \frac{m_i}{m}$
2. $\text{var}(Y_i) = n \frac{m_i}{m} \frac{m - m_i}{m}$
3. $\text{cov}(Y_i, Y_j) = -n \frac{m_i}{m} \frac{m_j}{m}$
4. $\text{cor}(Y_i, Y_j) = -\sqrt{\frac{m_i}{m - m_i} \frac{m_j}{m - m_j}}$

Comparing with our previous results, note that the means and correlations are the same, whether sampling with or without replacement. The variances and covariances are smaller when sampling without replacement, by a factor of the *finite population correction factor* $(m - n)/(m - 1)$

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. For the approximate multinomial distribution, we do not need to know m_i and m individually, but only in the ratio m_i/m.

Suppose that m_i depends on m and that $m_i/m \to p_i$ as $m \to \infty$ for $i \in \{1, 2, \ldots, k\}$. 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 $(p_1, p_2, \ldots, p_k)$.

Proof

Consider the second version of the hypergeometric probability density function. In the fraction, there are n factors in the denominator and n in the numerator. If we group the factors to form a product of n fractions, then each fraction in group i converges to p_i.

Examples and Applications

A population of 100 voters consists of 40 republicans, 35 democrats and 25 independents. A random sample of 10 voters is chosen. Find each of the following:

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

Answer

1. $\mathbb{P}(X = x, Y = y, Z = z) = \dfrac{\binom{40}{x}\binom{35}{y}\binom{25}{z}}{\binom{100}{10}}$ for $x,\ y,\ z \in \mathbb{N}$ with $x + y + z = 10$
2. $\mathbb{E}(X) = 4$, $\mathbb{E}(Y) = 3.5$, $\mathbb{E}(Z) = 2.5$
3. $\mathrm{var}(X) = 2.1818$, $\mathrm{var}(Y) = 2.0682$, $\mathrm{var}(Z) = 1.7045$
4. $\mathrm{cov}(X, Y) = -1.6346$, $\mathrm{cov}(X, Z) = -0.9091$, $\mathrm{cov}(Y, Z) = -0.7955$
5. 0.2474

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*.

In a bridge hand, find the probability density function of

1. The number of spades, number of hearts, and number of diamonds.
2. The number of spades and number of hearts.
3. The number of spades.
4. The number of red cards and the number of black cards.

Answer

Let X, Y, Z, U, and V denote the number of spades, hearts, diamonds, red cards, and black cards, respectively, in the hand.

1. $\mathbb{P}(X = x, Y = y, Z = z) = \dfrac{\binom{13}{x}\binom{13}{y}\binom{13}{z}\binom{13}{13-x-y-z}}{\binom{52}{13}}$ for $x,\ y,\ z \in \mathbb{N}$ with $x + y + z \leq 13$
2. $\mathbb{P}(X = x, Y = y) = \dfrac{\binom{13}{x}\binom{13}{y}\binom{26}{13-x-y}}{\binom{52}{13}}$ for $x,\ y \in \mathbb{N}$ with $x + y \leq 13$
3. $\mathbb{P}(X = x) = \dfrac{\binom{13}{x}\binom{39}{13-x}}{\binom{52}{13}}$ for $x \in \{0, 1, \ldots 13\}$
4. $\mathbb{P}(U = u, V = v) = \dfrac{\binom{26}{u}\binom{26}{v}}{\binom{52}{13}}$ for $u,\ v \in \mathbb{N}$ with $u + v = 13$

In a bridge hand, find each of the following:

1. The mean and variance of the number of spades.
2. The covariance and correlation between the number of spades and the number of hearts.
3. The mean and variance of the number of red cards.

Answer

Let X, Y, and U denote the number of spades, hearts, and red cards, respectively, in the hand.

1. $\mathbb{E}(X) = \frac{13}{4}$, $\mathrm{var}(X) = \frac{507}{272}$
2. $\mathrm{cov}(X, Y) = -\frac{169}{272}$
3. $\mathbb{E}(U) = \frac{13}{2}$, $\mathrm{var}(U) = \frac{169}{272}$

In a bridge hand, find each of the following:

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

Answer

Let X, Y and Z denote the number of spades, hearts, and diamonds respectively, in the hand.

1. $\mathbb{P}(X = x, Y = y, \mid Z = 4) = \dfrac{\binom{13}{x}\binom{13}{y}\binom{22}{9-x-y}}{\binom{48}{9}}$ for $x,\ y \in \mathbb{N}$ with $x + y \leq 9$
2. $\mathbb{P}(X = x \mid Y = 3, Z = 2) = \dfrac{\binom{13}{x}\binom{34}{8-x}}{\binom{47}{8}}$ for $x \in \{0, 1, \ldots, 8\}$

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

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

$$\frac{1913496}{2598960} \approx 0.736 \tag{12.3.12}$$

In the card experiment, set $n = 5$. Run the simulation 1000 times and 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.

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

$$\frac{32427298180}{635013559600} \approx 0.051 \tag{12.3.13}$$