Virtual Laboratories > 12. Finite Sampling Models > 1 2 3 4 5 6 7 8 9

2. The Hypergeometric Distribution

Basic Theory

Suppose that we have a **dichotomous** population D. That is, a population that consists of two types of objects, which we will refer to as type 1 and type 0. For example, we could have

- balls in an urn that are either *red* or *green*
- a batch of components that are either *good* or *defective*
- a population of people who are either *male* or *female*
- a population of animals that are either *tagged* or *untagged*

Let R denote the subset of D consisting of the type 1 objects, and suppose that $\#(D) = m$ and $\#(R) = r$. As in the basic sampling model, we sample n objects at random from D. In this section, our only concern is in the types of the objects, so let X_i denote the type of the i^{th} object chosen (1 or 0). The random vector of types is

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

Our main interest is the random variable Y that gives the number of type 1 objects in the sample. Note that Y is a counting variable, and thus like all counting variables, can be written as a sum of indicator variables, in this case the type variables:

$$Y = \sum_{i=1}^{n} X_i$$

We will assume initially that the sampling is without replacement, which is usually the realistic setting with dichotomous populations.

The Probability Density Function

Recall that since the sampling is without replacement, the unordered sample is uniformly distributed over the set of all combinations of size n chosen from D. This observation leads to a simple combinatorial derivation of the probability density function of Y.

1. Show that

$$\mathbb{P}(Y = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}, \quad k \in \{\max\{0, n - (m - r)\}, ..., \min\{n, r\}\}$$

This is known as the **hypergeometric distribution** with parameters m, r, and n.

2. Show the following alternative form of the hypergeometric probability density function in two ways:

combinatorially by treating the outcome as a permutation of size n chosen from the population of m balls, and *algebraically*, starting from the result in Exercise 1.

$$\mathbb{P}(Y = k) = \binom{n}{k} \frac{r^{(k)} (m - r)^{(n-k)}}{m^{(n)}}, \quad k \in \{\max\{0, n - (m - r)\}, ..., \min\{n, r\}\}$$

Recall our convention that $j^{(i)} = \binom{j}{i} = 0$ for $i > j$. With this convention, the formulas for the probability density function in Exercise 1 and Exercise 2 are correct for $k \in \{0, 1, ..., n\}$. We usually use this simpler set as the set of values for the hypergeometric distribution.

3. Let $v = \frac{(r+1)(n+1)}{m+1}$. Show that

a. $\mathbb{P}(Y = k) > \mathbb{P}(Y = k - 1)$ if and only if $k < v$.
b. The hypergeometric distribution is unimodal, at first increasing and then decreasing.
c. The mode occurs at $\lfloor v \rfloor$ if v is not an integer, and at v and $v - 1$ if v is an integer greater than 0.

4. In the ball and urn experiment, select sampling without replacement. Vary the parameters and note the shape of the probability density function. For selected values of the parameters, run the experiment 1000 times with an update frequency of 10 and watch the apparent convergence of the relative frequency function to the probability density function.

Moments

In the following exercises, we will derive the mean and variance of Y. The exchangeable property of the indicator variables, and properties of covariance and correlation will play a key role.

5. Show that $\mathbb{E}(X_i) = \frac{r}{m}$ for each i.

6. Show that $\mathbb{E}(Y) = n \frac{r}{m}$.

7. Show that $\operatorname{var}(X_i) = \frac{r}{m}\left(1 - \frac{r}{m}\right)$ for any i.

8. Show that for distinct i and j,

a. $\operatorname{cov}\left(X_i, X_j\right) = -\frac{r}{m}\left(1 - \frac{r}{m}\right)\frac{1}{m-1}$
b. $\operatorname{cor}\left(X_i, X_j\right) = -\frac{1}{m-1}$

Note from Exercise 8 that the event of a type 1 object on draw i and the event of a type 1 object on draw j are negatively correlated, but the correlation depends only on the population size and not on the number of type 1 objects. Note also that the correlation is perfect if $m = 2$. Think about these result intuitively.

9. Use the results of Exercise 7 and Exercise 8 to show that $\operatorname{var}(Y) = n \frac{r}{m}\left(1 - \frac{r}{m}\right)\frac{m-n}{m-1}$

Note that $\text{var}(Y) = 0$ if $r = 0$ or $r = m$ or $n = m$. Think about these results.

10. In the ball and urn experiment, select sampling without replacement. Vary the parameters and note the size and location of the mean/standard deviation bar. For selected values of the parameters, run the experiment 1000 times updating every 10 runs and watch the apparent convergence of the empirical moments to the true moments.

Sampling with Replacement

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

11. Show that $(X_1, X_2, ..., X_n)$ is a sequence of n Bernoulli trials with success parameter $\frac{r}{m}$.

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

12. Show that Y has the binomial distribution with parameters n and $\frac{r}{m}$:

$$\mathbb{P}(Y = k) = \binom{n}{k} \left(\frac{r}{m}\right)^k \left(1 - \frac{r}{m}\right)^{n-k}, \quad k \in \{0, 1, ..., n\}$$

13. Show that

a. $\mathbb{E}(Y) = n \frac{r}{m}$.

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

Note that for any values of the parameters, the mean of Y is the same, whether the sampling is with or without replacement. On the other hand, the variance of Y is smaller, by a factor of $\frac{m-n}{m-1}$, when the sampling is without replacement than with replacement. Think about these results. The factor $\frac{m-n}{m-1}$ is sometimes called the **finite population correction factor**.

14. In the ball and urn experiment, vary the parameters and switch between sampling without replacement and sampling with replacement. Note the difference between the graphs of the hypergeometric probability density function and the binomial probability density function. Note also the difference between the mean/standard deviation bars. For selected values of the parameters and for the two different sampling modes, run the simulation 1000 times, updating every 10 runs.

Convergence of the Hypergeometric Distribution to the Binomial

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 hypergeometric distribution should be well approximated by the binomial. The following exercise makes this observation precise. Practically, it is a valuable result, since the binomial distribution has fewer parameters. More specifically, we do not need to know the population size m and the number of type 1 objects r *individually*, but only in the *ratio* $\frac{r}{m}$.

15. Suppose that $r = r_m$ depends on m and that $\frac{r_m}{m} \to p$ as $m \to \infty$. Show that for fixed n, the hypergeometric probability density function with parameters m, r_m, and n converges to the binomial probability density function with parameters n and p. *Hint*: Use the representation in Exercise 2.

The type of convergence in the previous exercise is known as convergence in distribution.

16. In the ball and urn experiment, vary the parameters and switch between sampling without replacement and sampling with replacement. Note the difference between the graphs of the hypergeometric probability density function and the binomial probability density function. In particular, note the similarity when m is large and n small. For selected values of the parameters, and for both sampling modes, run the experiment 1000 times updating every 10 runs.

17. In the setting of Exercise 15, show that the mean and variance of the hypergeometric distribution converge to the mean and variance of the binomial distribution as $m \to \infty$.

Inferences in the Hypergeometric Model

In many real problems, the parameters r or m (or both) may be unknown. In this case we are interested in drawing inferences about the unknown parameters based on our observation of Y, the number of type 1 objects in the sample. We will assume initially that the sampling is without replacement, the realistic setting in most applications.

Estimation of r with m Known

Suppose that the size of the population m is known but that the number of type 1 objects r is unknown. This type of problem could arise, for example, if we had a batch of m manufactured items containing an unknown number r of defective items. It would be too costly and perhaps destructive to test all m items, so we might instead select n items at random and test those.

A simple estimator of r can be derived by hoping that the *sample* proportion of type 1 objects is close to the *population* proportion of type 1 objects. That is,

$$\frac{Y}{n} \approx \frac{r}{m} \text{ so } r \approx \frac{m}{n} Y$$

.

18. Show that $\mathbb{E}\left(\frac{m}{n} Y\right) = r$.

The result in the previous exercise means that $\frac{m}{n} Y$ is an **unbiased** estimator of r. Hence the variance is a measure of the quality of the estimator, in the mean square sense.

19. Show that $\operatorname{var}\left(\frac{m}{n} Y\right) = (m - r) \frac{r}{n} \frac{m-n}{m-1}$.

20. Show that for fixed m and r, $\operatorname{var}\left(\frac{m}{n} Y\right) \to 0$ as $n \uparrow m$.

Thus, the estimator improves as the sample size increases; this property is known as **consistency**.

21. In the ball and urn experiment, select sampling without replacement. For selected values of the parameters, run the experiment 100 times, updating after each run.

a. On each run, compare the true value of r with the estimated value.
b. Compute the average error and the average squared error over the 100 runs.
c. Compare the average squared error with the variance in Exercise 19.

Estimation of m with r Known

Suppose now that the number of type 1 objects r is known, but the population size m is unknown. As an example of this type of problem, suppose that we have a lake containing m fish where m is unknown. We capture r of the fish, tag them, and return them to the lake. Next we capture n of the fish and observe Y, the number of tagged fish in the sample. We wish to estimate m from this data. In this context, the estimation problem is sometimes called the **capture-recapture problem**.

22. Do you think that the main assumption of the sampling model, namely equally likely samples, would be satisfied for a real capture-recapture problem? Explain.

Once again, we can derive a simple estimate of m by hoping that the *sample* proportion of type 1 objects is close the *population* proportion of type 1 objects. That is,

$$\frac{Y}{n} \approx \frac{r}{m} \text{ so } m \approx \frac{n\,r}{Y}$$

Thus, our estimator of m is $\frac{n\,r}{Y}$ if $Y > 0$ and is undefined if $Y = 0$.

23. In the ball and urn experiment, select sampling without replacement. For selected values of the parameters, run the experiment 100 times, updating after each run.

a. On each run, compare the true value of m with the estimated value.
b. Compute the average error and the average squared error over the 100 runs.

24. Show that if $k > 0$ then $\frac{n\,r}{k}$ maximizes $\mathbb{P}(Y = k)$ as a function of m for fixed r and n. This means that $\frac{n\,r}{Y}$ is a maximum likelihood estimator of m.

25. Use Jensen's inequality to show that $\mathbb{E}\left(\frac{n\,r}{Y}\right) \ge m$.

Thus, the estimator is *biased* and tends to over-estimate m. Indeed, if $n \le m - r$, so that $\mathbb{P}(Y = 0) > 0$ then $\mathbb{E}\left(\frac{n\,r}{Y}\right) = \infty$.

For another approach to estimating m, see the section on Order Statistics.

Sampling with Replacement

Suppose now that the sampling is with replacement, even though this is unrealistic in most applications. In this case, Y has the binomial distribution with parameters n and $\frac{r}{m}$.

26. Show that

a. $\mathbb{E}\left(\frac{m}{n}Y\right) = r$.
b. $\operatorname{var}\left(\frac{m}{n}Y\right) = \frac{r\,(m-r)}{n}$.

Thus, the estimator of r with m known is still unbiased, but has larger mean square error. Thus, sampling without replacement works better, for any values of the parameters, than sampling with replacement.

27. In the ball and urn experiment, select sampling with replacement. For selected values of the parameters, run the experiment 100 times, updating after each run.

a. On each run, compare the true value of m with the estimated value.
b. Compute the average error and the average squared error over the 100 runs.

Examples and Applications

28. A batch of 100 computer chips contains 10 defective chips. Five chips are chosen at random, without replacement.

a. Compute the probability density function of the number of defective chips in the sample.
b. Compute the mean and variance of the number of defective chips in the sample
c. Find the probability that the sample contains at least one defective chip.

29. A club contains 50 members; 20 are men and 30 are women. A committee of 10 members is chosen at random.

a. Compute the probability density function of the number of women on the committee.
b. Give the mean and variance of the number of women on the committee.
c. Give the mean and variance of the number of men on the committee.
d. Find the probability that the committee members are all the same gender.

30. A small pond contains 1000 fish; 100 are tagged. Suppose that 20 fish are caught.

a. Compute the probability density function of the number of tagged fish in the sample.
b. Compute the mean and variance of the number of tagged fish in the sample.
c. Compute the probability that the sample contains at least 2 tagged fish.
d. Find the binomial approximation to the probability in (a).

31. Forty percent of the registered voters in a certain district prefer candidate *A*. Suppose that 10 voters are chosen at random.

a. Find the probability density function of the number of voters in the sample who prefer *A*.
b. Find the mean and variance of the number of voters in the sample who prefer *A*.
c. Find the probability that at least 5 voters in the sample prefer *A*.

32. Suppose that 10 memory chips are sampled at random and without replacement from a batch of 100 chips. The chips are tested and 2 are defective. Estimate the number of defective chips in the entire batch.

33. A voting district has 5000 registered voters. Suppose that 100 voters are selected at random and polled, and that 40 prefer candidate *A*. Estimate the number of voters in the district who prefer candidate *A*.

34. From a certain lake, 200 fish are caught, tagged and returned to the lake. Then 100 fish are caught and it turns out that 10 are tagged. Estimate the population of fish in the lake.

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**.

35. In a poker hand, find the probability density function, mean, and variance of the following random variables:

a. The number of spades
b. The number of aces

36. In a bridge hand, find the probability density function, mean, and variance of the following random variables:

a. The number of hearts
b. The number of honor cards (ace, king, queen, or jack).

The Randomized Urn

An interesting thing to do in almost any parametric probability model is to randomize one or more of the parameters. Done in the right way, this often leads to an interesting new parametric model, since the distribution of the randomized parameter will often itself belong to a parametric family. This is also the natural setting to apply Bayes' theorem.

In this section, we will randomize the number of type 1 objects in the basic hypergeometric model. Specifically, we assume that we have m objects in the population, as before. However, instead of a fixed number r of type 1 objects, we assume that each of the m objects in the population, independently of the others, is type 1 with probability p and type 0 with probability $1 - p$. We have eliminated one parameter, r, in favor of a new parameter p with values in the interval $[0, 1]$. Let U_i denote the type of the i^{th} object in the population, so that $\boldsymbol{U} = (U_1, U_2, ..., U_m)$ is a sequence of Bernoulli trials with success parameter p. Let $V = U_1 + U_2 + \cdots + U_m$ denote the number of type 1 objects in the population, so that V has the binomial distribution with parameters m and p.

As before, we sample n object from the population. Again we let X_i denote the type of the i^{th} object sampled, and we let $Y = X_1 + X_2 + \cdots + X_n$ denote the number of type 1 objects in the sample. We will consider sampling with and without replacement. In the first case, the sample size can be any positive integer, but in the second case, the sample size cannot exceed the population size. The key technique in the analysis of the randomized urn is to *condition on* V. If we know that $V = r$, then the model reduces to the model studied above: a population of size m with r type 1 objects, and a sample of size n.

37. Show that with either type of sampling,

$$\mathbb{P}(X_i = 1) = \mathbb{E}\left(\frac{V}{m}\right) = p$$

Thus, in either model, $\boldsymbol{X}$ is a sequence of identically distributed indicator variables. Ah, but what about dependence?

38. Suppose that the sampling is without replacement. Let $(x_1, x_2, ..., x_n) \in \{0, 1\}^n$ and let $y = x_1 + x_2 + \cdots + x_n$ Show that

$$\mathbb{P}(X_1 = x_1, X_2 = x_x, ..., X_n = x_n) = \mathbb{E}\left(\frac{V^{(y)}\,(m - V)^{(n-y)}}{m^{(n)}}\right) = p^y\,(1 - p)^{n-y}$$

a. Show the first equality by conditioning on V.
b. For the second equation, let $G(s, t) = \mathbb{E}\left(s^V\, t^{m-V}\right)$ and note that G is a probability generating function

of sorts.

c. Use the binomial theorem to show that $G(s,\, t) = (p\, s + (1 - p)\, t)^m$.

d. Let $G_{j,k}$ denote the partial derivative of G of order $j + k$, with j derivatives with respect to the first argument and k derivatives with respect to the second argument.

e. Use (b) to show that $G_{j,k}(1,\, 1) = \mathbb{E}\left(V^{(j)}\, (m - V)^{(k)}\right)$.

f. Use (c) to show that $G_{j,k}(1,\, 1) = m^{(j+k)}\, p^j\, (1 - p)^k$.

From the joint distribution in the previous exercise, we see that $\boldsymbol{X}$ is a sequence of Bernoulli trials with success parameter p, and hence Y has the binomial distribution with parameters n and p. We could also argue that $\boldsymbol{X}$ is a Bernoulli trials sequence directly, by noting that $\{X_1,\, X_2,\, ...,\, X_n\}$ is a randomly chosen subset of $\{U_1,\, U_2,\, ...,\, U_m\}$.

39. Suppose now that the sampling is with replacement. Again, let $(x_1,\, x_2,\, ...,\, x_n) \in \{0,\, 1\}^n$ and let $y = x_1 + x_2 + \cdots + x_n$ Show that

$$\mathbb{P}(X_1 = x_1,\, X_2 = x_{\text{x}},\, ...,\, X_n = x_n) = \mathbb{E}\left(\frac{V^y\, (m - V)^{n-y}}{m^n}\right)$$

A closed form expression for the joint distribution of $\boldsymbol{X}$, in terms of the parameters m, n, and p is not easy, but it is at least clear that the joint distribution will not be the same as the one when the sampling is without replacement. Thus, $\boldsymbol{X}$ is a dependent sequence. Note however that $\boldsymbol{X}$ is an exchangeable sequence, since the joint distribution is invariant under a permutation of the coordinates (this is a simple consequence of the fact that the joint distribution depends only on the sum y).

40. Note that

$$\mathbb{P}(Y = k) = \binom{n}{k} \mathbb{E}\left(\frac{V^k\, (m - V)^{n-k}}{m^n}\right), \quad k \in \{0,\, 1,\, ...,\, n\}$$

41. Let's compute the covariance and correlation of a pair of type variables when the sampling is with replacement. Suppose that i and j are distinct indices. Show that

a. $\mathbb{P}\left(X_i = 1,\, X_j = 1\right) = \mathbb{E}\left(\left(\frac{V}{m}\right)^2\right) = \frac{p\,(1-p)}{m} + p^2$

b. $\operatorname{cov}\left(X_i,\, X_j\right) = \frac{p\,(1-p)}{m}$.

c. $\operatorname{cor}\left(X_i,\, X_j\right) = \frac{1}{m}$.

42. Now we can get the mean and variance of Y. Show that

a. $\mathbb{E}(Y) = n\, p$

b. $\operatorname{var}(Y) = n\, p\, (1 - p)\, \frac{m+n-1}{m}$

Let's conclude with an interesting observation: For the randomized urn, $\boldsymbol{X}$ is a sequence of independent variables when the sampling is without replacement but a sequence of dependent variables when the sampling is with replacement--just the opposite of the situation for the deterministic urn with a fixed number of type 1 objects.

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