12.2: The Hypergeometric Distribution

Basic Theory

Dichotomous Populations

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*
- voters who are either *democrats* or *republicans*

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 ith object chosen (1 or 0). The random vector of types is

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \tag{12.2.1}$$

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:

$$Y = \sum_{i=1}^{n} X_i \tag{12.2.2}$$

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.

The probability density function of Y is given by

$$\mathbb{P}(Y = y) = \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}}, \quad y \in \{\max\{0, n-(m-r)\}, \ldots, \min\{n, r\}\} \tag{12.2.3}$$

Proof

Consider the unordered outcome, which is uniformly distributed on the set of combinations of size n chosen from the population of size m. The number of ways to select y type 1 objects from the r type 1 objects in the population is $\binom{r}{y}$. Similarly the number of ways to select the remaining $n - y$ type 0 objects from the $m - r$ type 0 objects in the population is $\binom{m-r}{n-y}$. Finally the number of ways to select the sample of size n from the population of size m is $\binom{m}{n}$.

This distribution defined by this probability density function is known as the *hypergeometric distribution* with parameters m, r, and n.

Another form of the probability density function of Y is

$$\mathbb{P}(Y = y) = \binom{n}{y}\frac{r^{(y)}(m-r)^{(n-y)}}{m^{(n)}}, \quad y \in \{\max\{0, n-(m-r)\}, \ldots, \min\{n, r\}\} \tag{12.2.4}$$

Combinatorial Proof

The combinatorial proof is much like the previous proof, except that we consider the ordered sample, which is uniformly distributed on the set of permutations of size n chosen from the population of m objects. The binomial coefficient $\binom{n}{y}$ is the number of ways to select the coordinates where the type 1 objects will go; $r^{(y)}$ is the number of ways to select an ordered sequence of y type 1 objects objects; and $(m-r)^{(n-y)}$ is the number of ways to select an ordered sequence of $n - y$ type 0 objects. Finally $m^{(n)}$ is the number of ways to select an ordered sequence of n objects from the population.

Algebraic Proof

The new form of the PDF can also be derived algebraically by starting with the previous form of the PDF. Use the formula $\binom{k}{j} = k^{(j)} / j!$ for each binomial coefficient, and then rearrange things a bit.

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

The hypergeometric distribution is unimodal. Let $v = \frac{(r+1)(n+1)}{m+2}$. Then

1. $\mathbb{P}(Y = y) > \mathbb{P}(Y = y - 1)$ if and only if $y < v$.
2. 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.

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 and compare the relative frequency function to the probability density function.

You may wonder about the rather exotic name *hypergeometric distribution*, which seems to have nothing to do with sampling from a dichotomous population. The name comes from a power series, which was studied by Leonhard Euler, Carl Friedrich Gauss, Bernhard Riemann, and others.

A (generalized) *hypergeometric series* is a power series

$$\sum_{k=0}^{\infty} a^k x^k \tag{12.2.5}$$

where $k \mapsto a_{k+1} / a_k$ is a rational function (that is, a ratio of polynomials).

Many of the basic power series studied in calculus are hypergeometric series, including the ordinary geometric series and the exponential series.

The probability generating function of the hypergeometric distribution is a hypergeometric series.

Proof

The PGF is $P(t) = \sum_{k=0}^{n} f(k) t^k$ where f is the hypergeometric PDF, given above. Simple algebra shows that

$$\frac{f(k+1)}{f(k)} = \frac{(r-k)(n-k)}{(k+1)(N-r-n+k+1)} \tag{12.2.6}$$

In addition, the hypergeometric distribution function can be expressed in terms of a hypergeometric *series*. These representations are not particularly helpful, so basically were stuck with the non-descriptive term for historical reasons.

Moments

Next 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.

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

Proof

Recall that X_i is an indicator variable with $\mathbb{P}(X_i = 1) = r / m$ for each i.

From the representation of Y as the sum of indicator variables, the expected value of Y is trivial to compute. But just for fun, we give the derivation from the probability density function as well.

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

Proof

This follows from the previous result and the additive property of expected value.

Proof from the definition

Using the hypergeometric PDF,

$$\mathbb{E}(Y) = \sum_{y=0}^{n} y \frac{\binom{r}{y}\binom{m-r}{n-y}}{\binom{m}{n}} \tag{12.2.7}$$

Note that the $y = 0$ term is 0. For the other terms, we can use the identity $y\binom{r}{y} = r\binom{r-1}{y-1}$ to get

$$\mathbb{E}(Y) = \frac{r}{\binom{m}{n}} \sum_{y=1}^{n} \binom{r-1}{y-1}\binom{m-r}{n-y} \tag{12.2.8}$$

But substituting $k = y - 1$ and using another fundamental identity,

$$\sum_{y=1}^{n} \binom{r-1}{y-1}\binom{m-r}{n-y} = \sum_{k=0}^{n-1} \binom{r-1}{k}\binom{m-r}{n-1-k} = \binom{m-1}{n-1} \tag{12.2.9}$$

So substituting and doing a bit of algebra gives $\mathbb{E}(Y) = n\frac{r}{m}$.

Next we turn to the variance of the hypergeometric distribution. For that, we will need not only the variances of the indicator variables, but their covariances as well.

$\text{var}(X_i) = \frac{r}{m}(1 - \frac{r}{m})$ for each i.

Proof

Again this follows because X_i is an indicator variable with $\mathbb{P}(X_i = 1) = r/m$ for each i.

For distinct i, j,

1. $\text{cov}(X_i, X_j) = -\frac{r}{m}(1 - \frac{r}{m})\frac{1}{m-1}$
2. $\text{cor}(X_i, X_j) = -\frac{1}{m-1}$

Proof

Note that $X_i X_j$ is an indicator variable that indicates the event that the ith and jth objects are both type 1. By the exchangeable property, $\mathbb{P}(X_i X_j = 1) = \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1 \mid X_i = 1) = \frac{r}{m}\frac{r-1}{m-1}$. Part (a) then follows from $\text{cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$. Part (b) follows from part (a) and the definition of correlation.

Note 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 $m = 2$, which must be the case.

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

Proof

This result follows from the previous results on the variance and covariance of the indicator variables. Recall that the variance of Y is the sum of $\text{cov}(X_i, X_j)$ over all i and j.

Note that $\text{var}(Y) = 0$ if $r = 0$ or $r = m$ or $n = m$, which must be true since Y is deterministic in each of these cases.

In the ball and urn experiment, select sampling without replacement. Vary the parameters and note the size and location of the mean $\pm$ standard deviation bar. For selected values of the parameters, run the experiment 1000 times and compare the empirical mean and standard deviation to the true mean and standard deviation.

Sampling with Replacement

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

$(X_1, X_2, \ldots, X_n)$ 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.

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

$$\mathbb{P}(Y = y) = \binom{n}{y} \left(\frac{r}{m}\right)^y \left(1 - \frac{r}{m}\right)^{n-y}, \quad y \in \{0, 1, \ldots, n\} \tag{12.2.10}$$

The mean and variance of Y are

1. $\mathbb{E}(Y) = n\frac{r}{m}$
2. $\text{var}(Y) = n\frac{r}{m}(1 - \frac{r}{m})$

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. It certainly makes sense that the variance of Y should be smaller when sampling without replacement, since each selection reduces the variablility in the population that remains. The factor $\frac{m-n}{m-1}$ is sometimes called the *finite population correction factor*.

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 $\pm$ standard deviation bars. For selected values of the parameters and for the two different sampling modes, run the simulation 1000 times.

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* r/m.

Suppose that $r_m \in \{0, 1, \ldots, m\}$ for each $m \in \mathbb{N}_+$ and that $r_m/m \to p \in [0, 1]$ as $m \to \infty$. Then 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 as $m \to \infty$

Proof

Consider the second version of the hypergeometric PDF above. In the fraction, note that there are n factors in the numerator and n in the denominator. Suppose we pair the factors to write the original fraction as the product of n fractions. The first y fractions have the form $\frac{r_m - i}{m - i}$ where i does not depend on m. Hence each of these fractions converge to p as $m \to \infty$. The remaining $n - y$ fractions have the form $\frac{m - r_m - j}{m - y - j}$, where again, j does not depend on m. Hence each of these fractions converges to $1 - p$ as $m \to \infty$.

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

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.

In the setting of the convergence result above, note 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 to test all m items (perhaps even destructive), 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,

$$\frac{Y}{n} \approx \frac{r}{m} \implies r \approx \frac{m}{n} Y \tag{12.2.11}$$

Thus, our estimator of r is $\frac{m}{n} Y$. This method of deriving an estimator is known as the method of moments.

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

Proof

This follows from the expected value of Y above, and the scale property of expected value.

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.

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

Proof

This follows from variance of Y above, and standard properties of variance.

For fixed m and r, $\text{var}\left(\frac{m}{n} Y\right) \downarrow 0$ as $n \uparrow m$.

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

In the ball and urn experiment, select sampling without replacement. For selected values of the parameters, run the experiment 100 times and note the estimate of r on each run.

1. Compute the average error and the average squared error over the 100 runs.
2. Compare the average squared error with the variance in mean square error given above.

Often we just want to estimate the *ratio* r/m (particularly if we don't know m either. In this case, the natural estimator is the sample proportion Y/n.

The estimator of $\frac{r}{m}$ has the following properties:

1. $\mathbb{E}\left(\frac{Y}{n}\right) = \frac{r}{m}$, so the estimator is unbiased.
2. $\text{var}\left(\frac{Y}{n}\right) = \frac{1}{n} \frac{r}{m} (1 - \frac{r}{m} \frac{m-n}{m-1})$
3. $\text{var}\left(\frac{Y}{n}\right) \downarrow 0$ as $n \uparrow m$ so the estimator is consistent.

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*.

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 use the method of moments to 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,

$$\frac{Y}{n} \approx \frac{r}{m} \implies m \approx \frac{nr}{Y} \tag{12.2.12}$$

Thus, our estimator of m is $\frac{nr}{Y}$ if $Y > 0$ and is ∞ if $Y = 0$.

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

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

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

$\mathbb{E}\left(\frac{nr}{Y}\right) \geq m$.

Proof

This result follows from Jensen's inequality since $y \mapsto \frac{nr}{y}$ is a convex function on $(0, \infty)$.

Thus, the estimator is *positivley biased* and tends to over-estimate m. Indeed, if $n \leq m - r$, so that $\mathbb{P}(Y = 0) > 0$ then $\mathbb{E}\left(\frac{nr}{Y}\right) = \infty$. For another approach to estimating the population size 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 $\frac{r}{m}$. The estimators of r with m known, $\frac{r}{m}$, and m with r known make sense, just as before, but have slightly different properties.

The estimator $\frac{m}{n}Y$ of r with m known satisfies

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

The estimator $\frac{1}{n}Y$ of $\frac{r}{m}$ satisfies

1. $\mathbb{E}\left(\frac{1}{n}Y\right) = \frac{r}{m}$
2. $\operatorname{var}\left(\frac{1}{n}Y\right) = \frac{1}{n}\frac{r}{m}\left(1 - \frac{r}{m}\right)$

Thus, the estimators are still unbiased and consistent, but have larger mean square error than before. Thus, sampling without replacement works better, for any values of the parameters, than sampling with replacement.

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

1. On each run, compare the true value of r with the estimated value.
2. Compute the average error and the average squared error over the 100 runs.

Examples and Applications

A batch of 100 computer chips contains 10 defective chips. Five chips are chosen at random, without replacement. Find each of the following:

1. The probability density function of the number of defective chips in the sample.
2. The mean and variance of the number of defective chips in the sample
3. The probability that the sample contains at least one defective chip.

Answer

Let Y denote the number of defective chips in the sample

1. $\mathbb{P}(Y = y) = \frac{\binom{10}{y}\binom{90}{5-y}}{\binom{100}{5}}, \quad y \in \{0, 1, 2, 3, 4, 5\}$
2. $\mathbb{E}(Y) = 0.5$, $\operatorname{var}(Y) = 0.432$
3. $\mathbb{P}(Y > 0) = 0.416$

A club contains 50 members; 20 are men and 30 are women. A committee of 10 members is chosen at random. Find each of the following:

1. The probability density function of the number of women on the committee.
2. The mean and variance of the number of women on the committee.
3. The mean and variance of the number of men on the committee.
4. The probability that the committee members are all the same gender.

Answer

Let Y denote the number of women, so that $Z = 10 - Y$ is the number of men.

1. $\mathbb{P}(Y = y) = \frac{\binom{30}{y}\binom{20}{10-y}}{\binom{50}{10}}, \quad y \in \{0, 1, \ldots, 10\}$
2. $\mathbb{E}(Y) = 6$, $\operatorname{var}(Y) = 1.959$
3. $\mathbb{E}(Z) = 4$, $\operatorname{var}(Z) = 1.959$

4. $\mathbb{P}(Y = 0) + \mathbb{P}(Y = 10) = 0.00294$

A small pond contains 1000 fish; 100 are tagged. Suppose that 20 fish are caught. Find each of the following:

1. The probability density function of the number of tagged fish in the sample.
2. The mean and variance of the number of tagged fish in the sample.
3. The probability that the sample contains at least 2 tagged fish.
4. The binomial approximation to the probability in (c).

Answer

Let Y denote the number of tagged fish in the sample

1. $\mathbb{P}(Y = y) = \dfrac{\binom{100}{y}\binom{900}{20-y}}{\binom{1000}{20}}, \quad y \in \{0, 1, \ldots, 20\}$
2. $\mathbb{E}(Y) = 2$, $\text{var}(Y) = \frac{196}{111}$
3. $\mathbb{P}(Y \geq 2) = 0.6108$
4. $\mathbb{P}(Y \geq 2) = 0.6083$

Forty percent of the registered voters in a certain district prefer candidate A. Suppose that 10 voters are chosen at random. Find each of the following:

1. The probability density function of the number of voters in the sample who prefer A.
2. The mean and variance of the number of voters in the sample who prefer A.
3. The probability that at least 5 voters in the sample prefer A.

Answer

1. $\mathbb{P}(Y = y) = \binom{10}{y}(0.4)^y(0.6)^{10-y}, \quad y \in \{0, 1, \ldots, 10\}$
2. $\mathbb{E}(Y) = 4$, $\text{var}(Y) = 2.4$
3. $\mathbb{P}(Y \geq 5) = 0.3669$

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.

Answer

20

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.

Answer

2000

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.

Answer

2000

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

1. The number of spades
2. The number of aces

Answer

Let U denote the number of spades and V the number of aces.

1. $\mathbb{P}(U = u) = \dfrac{\binom{13}{u}\binom{39}{5-u}}{\binom{52}{5}}, \quad u \in \{0, 1, \ldots, 5\}$, $\mathbb{E}(U) = \frac{5}{4}$, $\operatorname{var}(U) = \frac{235}{272}$
2. $\mathbb{P}(V = v) = \dfrac{\binom{4}{v}\binom{48}{5-v}}{\binom{52}{5}}, \quad v \in \{0, 1, 2, 3, 4\}$, $\mathbb{E}(V) = \frac{5}{13}$, $\operatorname{var}(V) = \frac{940}{2873}$

In a bridge hand, find each of the following:

1. The probability density function, mean, and variance of the number of hearts
2. The probability density function, mean, and variance of the number of honor cards (ace, king, queen, jack, or 10).
3. The probability that the hand has no honor cards. A hand of this kind is known as a Yarborough, in honor of Second Earl of Yarborough.

Answer

Let U denote the number of hearts and V the number of honor cards.

1. $\mathbb{P}(U = u) = \dfrac{\binom{13}{u}\binom{39}{13-u}}{\binom{52}{13}}, \quad u \in \{0, 1, \ldots, 13\}$, $\mathbb{E}(U) = \frac{13}{4}$, $\operatorname{var}(U) = \frac{507}{272}$
2. $\mathbb{P}(V = v) = \dfrac{\binom{20}{v}\binom{32}{13-v}}{\binom{52}{13}}, \quad v \in \{0, 1, \ldots, 13\}$, $\mathbb{E}(V) = 5$, $\operatorname{var}(V) = 2.353$
3. $\frac{5394}{9\,860\,459} \approx 0.000547$

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 ith object in the population, so that $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)$ is a sequence of Bernoulli trials with success parameter p. Let $V = \sum_{i=1}^m U_i$ 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 ith object sampled, and we let $Y = \sum_{i=1}^n X_i$ 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.

With either type of sampling, $\mathbb{P}(X_i = 1) = p$

Proof

$\mathbb{P}(X_i = 1) = \mathbb{E}[\mathbb{P}(X_i = 1 \mid V)] = \mathbb{E}(V/m) = p$

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

Suppose that the sampling is without replacement. Let $(x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ and let $y = \sum_{i=1}^n x_i$. Then

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p^y (1 - p)^{n-y} \tag{12.2.13}$$

Proof

Conditioning on V gives

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{E}[\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid V)] = \mathbb{E}\left[\frac{V^{(y)} (m - V)^{(n-y)}}{m^{(n)}}\right] \tag{12.2.14}$$

Now let $G(s, t) = \mathbb{E}(s^V t^{m-V})$. Note that G is a is a probability generating function of sorts. From the binomial theorem, $G(s, t) = [ps + (1 - p)t]^m$. 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. From the definition of G, $G_{j,k}(1, 1) = \mathbb{E}[V^{(j)} (m - V)^{(k)}]$. But from the binomial representation, $G_{j,k}(1, 1) = m^{j+k} p^j (1 - p)^k$

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 $\{X_1, X_2, \ldots, X_n\}$ is a randomly chosen subset of $\{U_1, U_2, \ldots, U_m\}$.

Suppose now that the sampling is with replacement. Again, let $(x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ and let $y = \sum_{i=1}^n x_i$. Then

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{E}\left[\frac{V^y (m - V)^{n-y}}{m^n}\right] \tag{12.2.15}$$

Proof

The result follows as before by conditioning on V:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{E}\left[\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid V)\right] = \mathbb{E}\left[\frac{V^y (m - V)^{n-y}}{m^n}\right] \tag{12.2.16}$$

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. In particular, $\boldsymbol{X}$ 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).

The probability density function of Y is given by

$$\mathbb{P}(Y = y) = \binom{n}{y} \mathbb{E}\left[\frac{V^y (m - V)^{n-y}}{m^n}\right], \quad y \in \{0, 1, \ldots, n\} \tag{12.2.17}$$

Suppose that i and j are distinct indices. The covariance and correlation of (X_i, X_j) are

1. $\operatorname{cov}(X_i, X_j) = \frac{p(1-p)}{m}$
2. $\operatorname{cor}(X_i, X_j) = \frac{1}{m}$

Proof

Conditioning on V once again we have $\mathbb{P}(X_i = 1, X_j = 1) = \mathbb{E}\left[\left(\frac{V}{m}\right)^2\right] = \frac{p(1-p)}{m} + p^2$. The results now follow from standard formulas for covariance and correlation.

The mean and variance of Y are

1. $\mathbb{E}(Y) = np$
2. $\operatorname{var}(Y) = np(1-p)\frac{m+n-1}{m}$

Proof

Part (a) follows from the distribution of the indicator variables above, and the additive property of expected value. Part (b) follows from the previous result on covariance. Recall again that the variance of Y is the sum of $\operatorname{cov}(X_i, X_j)$ over all i and j.

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.
