The Hypergeometric Distribution

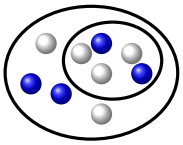
Mathias Winther Madsen

January 14, 2015

Problem 1. Among the 100 applicants for a job, 20 have the necessary qualifications. How many of the 100 applicants do you need to interview in order to be 95% sure to meet at least 10 qualified ones?

1 Sampling Without Replacement

A bag contains B blue and W white marbles, and that I grab a handful of marbles out of this bag, all at once. What is the probability that my sample contains exactly b blue and w white marbles, given that the sample size $b + w$ is held fixed?



To solve this problem, we need to count the total number of such samples and then count how many of them that have the required numbers of blue and white marbles. Since there was a total of $B + W$ marbles in the bag, and that we grabbed $b + w$ of them, the total number of possible samples was

$$\left(\begin{array}{c} B+W \\ b+w \end{array} \right).$$

We then want to count how many of these samples contain exactly b blue and w white marbles.

The number of ways we can choose b blue marbles from a pool B is counted by the binomial coefficient $C(B, b)$. The number of ways we can choose w white marbles from a pool of W is simarly counted by $C(W, w)$. By combining these two selections in all possible ways, we thus find that there is

$$\left(\begin{array}{c} B \\ b \end{array} \right) \left(\begin{array}{c} W \\ w \end{array} \right)$$

ways of selecting a sample of size $b + w$ containing b blue and w white marbles. The probability of drawing such a "successful" sample is consequently

$$\frac{\left(\begin{array}{c} B \\ b \end{array} \right) \left(\begin{array}{c} W \\ w \end{array} \right)}{\left(\begin{array}{c} B+W \\ b+w \end{array} \right)}.$$

For instance, if the bag contains $B + W = 100$ marbles, with $B = W = 50$, the probability of observing $b = 4$ blue and $w = 6$ white marbles is

$$\frac{\left(\begin{array}{c} 50 \\ 4 \end{array} \right) \left(\begin{array}{c} 50 \\ 6 \end{array} \right)}{\left(\begin{array}{c} 100 \\ 10 \end{array} \right)} = \frac{1727250}{8170019} \approx 0.211.$$

For $b = w = 5$, this probability is about 0.259, and for $b = 10$ and $w = 0$, 0.001.

The probability that we would randomly draw a sample that faithfully represents the proportions in population is thus substantial, but there is also a reasonably high probability that the composition of the sample will deviate slightly from the composition of the population.

The distribution that describe these combinatorical effects is called the **hypergeometric distribution**. Like the binomial distribution, the hypergeometric distribution counts the number of successes in a sample of a fixed size. Unlike the binomial distribution, however, it assumes that the population has a finite size, and that we sample without replacement.

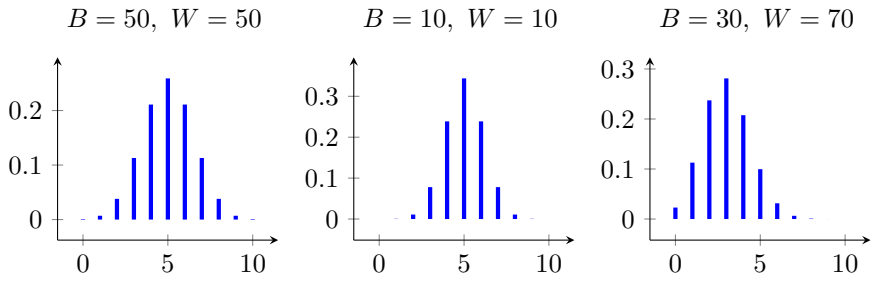


Figure 1: The probability of drawing b blue marbles when randomly selecting $b + w = 10$ marbles from a bag of B blue and W white marbles.

2 Expected Number of Successes

The parameterization of the hypergeometric distribution in terms of B, W, b, and w is intuitive, but not very convenient for all mathematical purposes. A more common way of parametrizing the distribution is therefore in terms of the following parameters:

1. $T = B + W$, the **total pupulation size**;
2. $t = b + w$, the **total sample size**;
3. $S = B$, the number of **"successes" in the population**;
4. $s = b$, the number of **"successes" in the sample**.

Using this parametrization, we have the following point probabilities:

$$\Pr(s \mid T, S, t) \;=\; \frac{\binom{S}{s}\binom{T-S}{t-s}}{\binom{T}{t}}.$$

In the first of the marbles examples discussed above, for instance, we had $T = 100$, $S = 50$, $t = 10$, and $s = 4$. We will consider S, T, and t as parameters of the distribution of the random variable s.

As you might guess, the expected value of the variable s is

$$E[s] \;=\; \frac{S}{T}\,t.$$

That is, the proportion of successes in the population scaled up to match the sample size. In order to prove this formula, we will make use of the following downscaling identity for binomial coefficients:

$$\binom{n}{k} \;=\; \binom{n-1}{k-1}\frac{n}{k}.$$

Theorem 2. *The mean of the hypergeometric distribution is* $E[s] = tS/T$.

Proof. The mean is an average over the $S+1$ possible values of s:

$$E[s] \;=\; \sum_{s=0}^{S} \Pr\left(s \,|\, T, S, t\right) s \;=\; \sum_{s=1}^{S} \Pr\left(s \,|\, T, S, t\right) s.$$

By applying the downscaling identity to the binomial coefficients $C(S,s)$ and $C(T,t)$ which occur in the hypergeometric probabilities, we can rewrite this as

$$\sum_{s=1}^{S} \frac{\left(\begin{array}{c} S \\ s \end{array}\right)\left(\begin{array}{c} T-S \\ t-s \end{array}\right)}{\left(\begin{array}{c} T \\ t \end{array}\right)} s \;=\; \sum_{s=1}^{S} \frac{\left(\begin{array}{c} S-1 \\ s-1 \end{array}\right)\left(\begin{array}{c} T-S \\ t-s \end{array}\right)}{\left(\begin{array}{c} T-1 \\ t-1 \end{array}\right)} \frac{S/s}{T/t} s.$$

After cancelling s/s and pulling tS/T out of the summation, we can make a change of summation variable to rewrite this once again as

$$\frac{tS}{T} \times \sum_{s=1}^{S} \frac{\left(\begin{array}{c} S-1 \\ s-1 \end{array}\right)\left(\begin{array}{c} T-S \\ t-s \end{array}\right)}{\left(\begin{array}{c} T-1 \\ t-1 \end{array}\right)} \;=\; \frac{tS}{T} \times \sum_{s=0}^{S-1} \frac{\left(\begin{array}{c} S-1 \\ s \end{array}\right)\left(\begin{array}{c} T-S \\ t-1-s \end{array}\right)}{\left(\begin{array}{c} T-1 \\ t-1 \end{array}\right)}.$$

Now, an inspection of the terms in the sum shows that these are in fact the probabilities of a hypergeometric distribution with parameters $T-1$, $S-1$, and $t-1$. The sum consequently evaluates to 1, and $E[s] = tS/T$. □

Note that this derivation incidentally also shows that decreasing all the parameters of a hypergeometric distribution by one corresponds to adjusting all of the probabilities by the factor $(s/t)/(S/T)$.

3 Statistical Applications

The hypergeometric distribution can be used for a number of interesting things in statistics. One such application is **Fisher's exact test**, a statistical test that measures whether there is a significant statistical dependence between two binary variables.

As an example, let's look the survival rates at the Titanic. There were $T = 2201$ people on the ship when it sank, $S = 771$ of whom survived. Of the $T = 2201$ people on the ship, $t = 325$ held a first-class ticket. We are interested in making some hypothetical statements about s, the number of people with a first-class ticket who survived, in order to compare those deductions to reality.

If having a first-class ticket made no difference as to whether you survived or not, then whe should expect the $S = 771$ survivors to be randomly scattered across the two groups, the people with and without first-class tickets. This would mean that we would expect to find about

$$E[s] \;=\; \frac{tS}{T} \;=\; 104.98,$$

survivers among our sample of $t = 325$ people. To illustrate how this tendency would work in practice, here are a few samples from a hypergeometric distribution with parameters $T = 2201$, $S = 771$ and $t = 325$:

$$s \;=\; 89, 104, 105, 92, 110, 93, 102, 103, 104, 103, 103, 93, 103, 115, 96.$$

As you can see, these values cluster quite reliably around the mean, 105. The sample average is $\bar{s} = 101.00$, and the largest deviation from the mean in this sample is roughly $|89 - 104.98| = 15.98$. So if holding a first-class ticket really made no difference to your changes of survival, we should thus expect the survival statistics to give rise to 2×2 contingency tables like

	Survived	Died	Totals
First-class ticket	100	225	325
Other ticket	611	1215	1876
Totals	711	1490	2201

By contrast, we would be more surprised if we were to see a disproportionately large number of survivors among the first-class ticket holders, as in

	Survived	Died	Totals
First-class ticket	150	175	325
Other ticket	561	1315	1876
Totals	711	1490	2201

These suspicions are confirmed by computing the two point probabilities:

$$\begin{aligned} \Pr\{s = 100 \,|\, T = 2201, S = 711, t = 325\} &\approx 4.2 \times 10^{-2}. \\ \Pr\{s = 150 \,|\, T = 2201, S = 711, t = 325\} &\approx 5.7 \times 10^{-9}. \end{aligned}$$

A value of $s = 100$ is thus about ten million times more probable than a value of $s = 150$ under the independence assumption. This indicates that while a deviation of about $|s - E[s]| = 5$ is reasonably likely to happen by chance, a deviation of $|s - E[s]| = 45$ is not. This gives you a sense of roughly how much s varies around its mean, $E[s] \approx 105$.

Now that you have a sense of which 2×2 contingency tables we might plausibly see under the independence assumption, let's compare that hypothetical behavior with the actual numbers:

	Survived	Died	Totals
First-class ticket	203	122	325
Other ticket	508	1368	1876
Totals	711	1490	2201

Obviously, this is way outside the bounds of the reasonable under the inde-

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested.

⋮

Our experiment consists in mixing eight cups of tea, four in one way and four in another, and presenting them to the subject for judgment in a random order.

⋮

At best, the subject can judge rightly with every cup and, knowing that 4 are of each kind, this amounts to choosing, out of the 70 sets of 4 which might be chosen, that particular one which is correct. A subject without any faculty of discrimination would in fact divide the 8 cups correctly into two sets of 4 in one trial out of 70, or, more properly, with a frequency which would approach 1 in 70 more and more nearly the more often the test were repeated.

Ronald Fisher: *The Design of Experiments* (1935), Chapter II.

pendence assumption. In fact, the probability of the event

$$G \;=\; \{s \geq 203\}$$

is indistinguishable from 0 with the parameter values $T = 2201$, $S = 711$, and $t = 325$. We can thus conclude that having a first-class ticket is extremely unlikely to be independent of your chances of survival.

By comparison, the set $G = \{s \geq 117\}$ has the more moderate probability of about 5.4%. It would thus not be completely unexpected to see $s = 117$ survivers among the $t = 325$ people with first class tickets, even under an independence assumption. Had we therefore observed the value $s = 117$, we could not have been quite as confident about our conclusion: 117 is larger than 105, but not quite enough larger to warrant any strong conclusions.