Hypergeometric Probability Distribution

Example problem:
Suppose 30 people have been summoned for jury selection, and that 12 people will be chosen entirely at random (not how the real process works!). Also, suppose that there are 17 candidates that are less than 40 years old, and 13 candidates that are at least 40 years old. What is the probability that exactly 5 of the candidates chosen are less than 40 years old?

Solution:
We must choose 5 “younger” candidates from the 17 available, and 7 “older” candidates from the 13 available. This will be a total of 12 jurors.

Since each person is equally likely to be chosen, the probability that exactly 5 “younger” candidates are chosen is

\[ P(5) = \frac{\binom{17}{5} \cdot \binom{13}{7}}{\binom{30}{12}} \]

\[ \approx 28.9\% \]

The Distribution
This is an example of the hypergeometric distribution:

- there are \( n \) possible outcomes. This is sometimes called the “population size”.
- there are \( a \) outcomes which are classified as “successes” (and therefore \( n - a \) “failures”)
- there are \( r \) trials. This is sometimes called the “sample size”.
- the trials are dependent
- the random variable \( X \) measures the number of successes

Then the probability that exactly \( k \) successes occur in the \( r \) trials is

\[ P(X = k) = \frac{\binom{a}{k} \cdot \binom{n - a}{r - k}}{\binom{n}{r}} \]

In the jury example above, we have the following parameters:

\( n = 30 \)
\( a = 17 \)
\( r = 12 \)
\( k = 5 \)
Expected Value

The expected value for a hypergeometric distribution is the *number of trials* multiplied by the *proportion of the population that is successes*:

\[ E(X) = \frac{ra}{n} \]

**Example 1: Drawing 2 Face Cards**

Suppose you draw 5 cards from a standard, shuffled deck of 52 cards. What is the probability that you draw *exactly 2 face cards*? What is the *expected number* of face cards?

**Solution**

This is a hypergeometric distribution with the following values:

- \( n = 52 \) possible outcomes
- \( a = 12 \) successes (face cards)
- \( r = 5 \) trials (cards drawn)
- \( k = 2 \) required successes (face cards)

\[
P(X = k) = \binom{a}{k} \frac{\binom{n-a}{r-k}}{\binom{n}{r}}
\]

\[
P(X = 2) = \binom{12}{2} \frac{\binom{52-12}{5-2}}{\binom{52}{5}} = \binom{12}{2} \frac{\binom{40}{3}}{\binom{52}{5}}
\]

\[
P(X = 2) \approx 0.2509
\]

The probability of getting exactly two face cards is about 25%.

The expected number of face cards in a hand of 5 cards is

\[
E(X) = \frac{ra}{n} = \frac{5 \times 12}{52} = \frac{15}{13} \approx 1.15
\]
Example 2: Gender split for hiring

There are 85 people who interview for 4 data science positions at a prestigious company. The company ranks the applicants and decides there are 12 candidates who are all equally suited for the 4 positions. Since they are concerned about unfair biases in the selection process, the company decides to choose 4 people at random from the 12 best candidates.

7 of the 12 candidates identify as female, and 5 of the 12 candidates identify as male. What is the probability that there will be exactly 2 people hired who identify with each gender? What is the expected number hired by gender?

Solution

This is a hypergeometric distribution with the following values:

- \( n = 12 \) (total number of candidates)
- \( a = 7 \) (candidates identifying as female)
- \( r = 4 \) (required number of candidates)
- \( k = 2 \) (required number of candidates identifying as female)

Note that we have arbitrarily selected candidates who identify as female as the “success” category; the same procedure works by selecting candidates who identify as male as the “success” category.

\[
P(X = k) = \frac{\binom{a}{k} \binom{n-a}{r-k}}{\binom{n}{r}}
\]

\[
P(X = 2) = \frac{\binom{7}{2} \binom{12-7}{4-2}}{\binom{12}{4}}
\]

\[
P(X = 2) = \frac{\binom{7}{2} \binom{5}{2}}{\binom{12}{4}}
\]

\[
P(X = 2) \approx 0.424
\]

There is about a 42.4% probability that there will be 2 candidates hired who identify with each gender.

\[
E(X) = \frac{ra}{n} = \frac{4 \times 7}{12} = \frac{7}{3} \approx 2.33
\]

The expected number of candidates hired who identify as female is about 2.33, and the expected number of candidates hired who identify as male is about \( 4 - 2.33 = 1.67 \).
Example 3: M:tG
Suppose you have a Magic: the Gathering deck of 60 cards, of which 22 are lands and 38 are non-lands. Make a probability distribution table for lands drawn in the opening hand of 7 cards. Use the table to calculate the probability of drawing 2 or 3 lands in the opening hand.

Solution
This is a hypergeometric distribution, with the following values (counting land cards as successes):

\[ n = 60 \] (total number of cards)
\[ a = 22 \] (land cards)
\[ r = 7 \] (cards drawn)

We need to calculate \( P(X = k) \) for each \( k \in \{0,1,2,\ldots,7\} \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>( P(X = k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.27%</td>
</tr>
<tr>
<td>1</td>
<td>15.73%</td>
</tr>
<tr>
<td>2</td>
<td>30.02%</td>
</tr>
<tr>
<td>3</td>
<td>29.43%</td>
</tr>
<tr>
<td>4</td>
<td>15.98%</td>
</tr>
<tr>
<td>5</td>
<td>4.79%</td>
</tr>
<tr>
<td>6</td>
<td>0.73%</td>
</tr>
<tr>
<td>7</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

The probability of drawing an opening hand with 2 or 3 land cards is about \( 30.02\% + 29.43\% = 59.45\% \).

Example 4: \( k \) can’t exceed \( r \)
There are 15 players available for a tournament. 5 players will be selected at random to participate. Only 3 players are experienced players. What is the probability that at least 2 players selected will be experienced?

Solution
Even though there are 5 trials \( (r = 5) \) it is not possible to have more than 3 successes (since \( a = 3 \)).

\[
P(X \geq 2) = P(X = 2) + P(X = 3)
\]

\[
= \frac{\binom{3}{2} \binom{15 - 3}{5 - 2}}{\binom{15}{5}} + \frac{\binom{3}{3} \binom{15 - 3}{5 - 3}}{\binom{15}{5}}
\]

\[
= \frac{\binom{3}{2} \binom{12}{3}}{\binom{15}{5}} + \frac{\binom{3}{3} \binom{12}{2}}{\binom{15}{5}}
\]

\[
= \frac{660}{3003} + \frac{66}{3003}
\]
\[
\frac{22}{91} = 24.2\%
\]

It’s not possible to have \( k = 4 \), for example. If you were to try to substitute the value \( k = 4 \) into the probability formula, you would have the combination \( \binom{3}{4} \), which has no value. Instead, we say that \( P(X = 4) = 0 \) in this situation.

**Hypergeometric Distribution vs. Binomial Distribution**

Both distributions have a fixed number of trials.

The trials in hypergeometric distributions are *dependent*, and the trials in binomial distributions are *independent*.

In a hypergeometric distribution there is a fixed number of possible successes available, and they’re “used up” as trials occur. In a binomial distribution the successes are not “used up”. We sometimes call this drawing “without replacement” (hypergeometric) versus “with replacement” (binomial).