Hypergeometric Random Variables

Now we consider a probability distribution that can be considered to arise from repeated Bernoulli trials where the probability of success **is not** constant.

Hypergeometric Random Variables

Now we consider a probability distribution that can be considered to arise from repeated Bernoulli trials where the probability of success **is not** constant.

The simplest way to describe this distribution is the **urn experiment**

Hypergeometric Random Variables

Now we consider a probability distribution that can be considered to arise from repeated Bernoulli trials where the probability of success **is not** constant.

The simplest way to describe this distribution is the **urn experiment**

An urn initially contains a certain number r of red chips and a certain number b of black chips.

Hypergeometric Random Variables

Now we consider a probability distribution that can be considered to arise from repeated Bernoulli trials where the probability of success **is not** constant.

The simplest way to describe this distribution is the **urn experiment**

An urn initially contains a certain number r of red chips and a certain number b of black chips.

The experiment consists of drawing some predetermined number n of chips from the urn, **without replacement**, and noting the number of red and black chips.

Hypergeometric Random Variables

Now we consider a probability distribution that can be considered to arise from repeated Bernoulli trials where the probability of success **is not** constant.

The simplest way to describe this distribution is the **urn experiment**

An urn initially contains a certain number r of red chips and a certain number b of black chips.

The experiment consists of drawing some predetermined number n of chips from the urn, **without replacement**, and noting the number of red and black chips.

The random variable associated with the experiment is usually defined to be the number of red chips drawn.

Hypergeometric Random Variables

Had we replaced the chip after each draw, a bit of thought should convince you that the number of red chips drawn should have a binomial distribution, because we conduct a fixed number n of Bernoulli trials each with probability of success

$$p = \frac{r}{r+b}$$

Hypergeometric Random Variables

Had we replaced the chip after each draw, a bit of thought should convince you that the number of red chips drawn should have a binomial distribution, because we conduct a fixed number n of Bernoulli trials each with probability of success

$$p = \frac{r}{r+b}$$

The slight modification in the experiment that produces the hypergeometric distribution instead of the binomial is that we **do not** replace each chip after it is drawn, but conduct the next draw from whatever chips are left in the urn.

Hypergeometric Random Variables

Had we replaced the chip after each draw, a bit of thought should convince you that the number of red chips drawn should have a binomial distribution, because we conduct a fixed number n of Bernoulli trials each with probability of success

$$p = \frac{r}{r+b}$$

The slight modification in the experiment that produces the hypergeometric distribution instead of the binomial is that we **do not** replace each chip after it is drawn, but conduct the next draw from whatever chips are left in the urn.

Because we do not replace chips as we draw them, the probability of drawing a red chip does not remain the same on successive draws.

Hypergeometric Random Variables

As it turns out, the probability of a red chip on the second draw depends on the outcome of the first draw.

If the first draw results in a red chip, the probability of a red chip on the second draw is:

$$p = \frac{r-1}{r+b-1}$$

Hypergeometric Random Variables

As it turns out, the probability of a red chip on the second draw depends on the outcome of the first draw.

If the first draw results in a red chip, the probability of a red chip on the second draw is:

$$p = \frac{r-1}{r+b-1}$$

However, if the first chip drawn is black, the probability of a red chip on the second draw is:

$$p = \frac{r}{r+b-1}$$

Hypergeometric Random Variables

Because we violate both the assumption of independence of successive trials, and the assumption of constant probability of success, this is not a binomial experiment.

Hypergeometric Random Variables

Because we violate both the assumption of independence of successive trials, and the assumption of constant probability of success, this is not a binomial experiment.

The distribution that results from this experiment is called the **hypergeometric** distribution.

Hypergeometric Random Variables

The author's notation for the hypergeometric distribution is:

$$P(X = x) = h(x; n, M, N)$$

where, in the parlance of the urn experiment,

- N represents the number of chips in the urn at the start
- M represents the number of red chips
- n represents the number of chips drawn (and not replaced) during the experiment
- x represents the number of red chips drawn

Hypergeometric Random Variables

The author's notation for the hypergeometric distribution is:

$$P(X = x) = h(x; n, M, N)$$

where, in the parlance of the urn experiment,

- N represents the number of chips in the urn at the start
- M represents the number of red chips
- n represents the number of chips drawn (and not replaced) during the experiment
- x represents the number of red chips drawn

As we see, the hypergeometric is a bit more complicated than the others we have studied.

Hypergeometric Random Variables

The probability mass function for the hypergeometric distribution is:

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

with some restrictions on the values that x and n can take:

- x cannot be smaller than zero
- x cannot be larger than M, the initial number of red chips
- x cannot be larger than n, the number of chips drawn
- n cannot be larger than N, the number of chips initially in the urn

Hypergeometric Random Variables

The probability mass function for the hypergeometric distribution is:

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

with some restrictions on the values that x and n can take:

- x cannot be smaller than zero
- x cannot be larger than M, the initial number of red chips
- x cannot be larger than n, the number of chips drawn
- n cannot be larger than N, the number of chips initially in the urn

Hypergeometric Random Variables

Because of the complicated nature of the probability mass function, there is no simple formula for the cumulative distribution function, defined by $F(x) = P(X \leq x)$.

Hypergeometric Random Variables

Because of the complicated nature of the probability mass function, there is no simple formula for the cumulative distribution function, defined by $F(x) = P(X \leq x)$.

With a bit of tedious algebra, one can show that the expected value and variance of a hypergeometric random variable are:

$$E(X) = n \cdot \frac{M}{N} \quad \text{and} \quad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

Hypergeometric Random Variables

Because of the complicated nature of the probability mass function, there is no simple formula for the cumulative distribution function, defined by $F(x) = P(X \leq x)$.

With a bit of tedious algebra, one can show that the expected value and variance of a hypergeometric random variable are:

$$E(X) = n \cdot \frac{M}{N} \quad \text{and} \quad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

Note that if we let $p = M/N$ and we let N become large while p remains fixed, we end up with the mean and variance of a binomial random variable.

Hypergeometric Random Variables

Once again, there is not universal agreement on the $h(x; n, M, N)$ notation.

Hypergeometric Random Variables

Once again, there is not universal agreement on the $h(x; n, M, N)$ notation.

The R implementation of the hypergeometric distribution as expected is `dhyper`, but the parameters are slightly different.

Hypergeometric Random Variables

Once again, there is not universal agreement on the $h(x; n, M, N)$ notation.

The R implementation of the hypergeometric distribution as expected is `dhyper`, but the parameters are slightly different.

In terms of the author's notation,

$$P(X = x) = h(x; n, M, N) = \texttt{dhyper(x,M,N-M,n)}$$

Hypergeometric Random Variables

Example: Keno

In the lottery game called Keno, 20 of the 80 numbers from 1 to 80 are randomly selected. Prior to the selection, players fill out a card specifying up to 10 numbers they think will be chosen.

Hypergeometric Random Variables

Example: Keno

In the lottery game called Keno, 20 of the 80 numbers from 1 to 80 are randomly selected. Prior to the selection, players fill out a card specifying up to 10 numbers they think will be chosen.

If the player chooses 10 numbers, think of the 80 numbers as divided into 10 "winning" numbers, and 70 "losing" numbers.

Hypergeometric Random Variables

Example: Keno

In the lottery game called Keno, 20 of the 80 numbers from 1 to 80 are randomly selected. Prior to the selection, players fill out a card specifying up to 10 numbers they think will be chosen.

If the player chooses 10 numbers, think of the 80 numbers as divided into 10 "winning" numbers, and 70 "losing" numbers.

The experiment is then hypergeometric with 20 numbers chosen without replacement.

Hypergeometric Random Variables

Example: Keno

In the lottery game called Keno, 20 of the 80 numbers from 1 to 80 are randomly selected. Prior to the selection, players fill out a card specifying up to 10 numbers they think will be chosen.

If the player chooses 10 numbers, think of the 80 numbers as divided into 10 "winning" numbers, and 70 "losing" numbers.

The experiment is then hypergeometric with 20 numbers chosen without replacement.

The probablility of getting 5 hits out of 10 numbers chosen is:

$$P(X=5) = h(5;20,10,80) = \texttt{dhyper(5,10,80-10,20)} = .051$$

Hypergeometric Random Variables

The hypergeometric can also be visualized by a tree diagram.