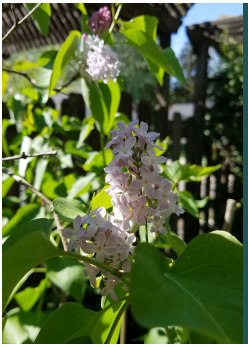


Module C: Understanding the Data-Generating Process

Slide Deck C4:

# Hypergeometric Distributions

*The section in which we learn about a fourth named discrete distributions. A random variable that follows a Hypergeometric distribution will model the number of successes out of a number of trials, but without replacement.*



## Today's Objectives

By the end of this slidedeck, you should

- determine what types of random variables follow a Hypergeometric distribution using its statistical definition
- calculate the usual two important population parameters:
  - expected value
  - variance
- identify the three parameters defining the Hypergeometric distribution
- calculate probabilities associated with a Hypergeometric random variable

## Definition of a Hypergeometric Experiment

**Definition**

The **Hypergeometric distribution** describes the probability of obtaining $x$ successes in $k$ draws (trials), *without* replacement, from a finite population that contains exactly $m$ successes and $n$ failures.

**Examples**

- number of hearts drawn from a deck without replacement
- number of Sophomores counted in the library (without re-counting)
- number of parolees that return to the Hill Correctional Center in Galesburg

The difference between a Binomial and a Hypergeometric distribution is the size of the population (infinite or finite) and whether things can be selected multiple times (yes or no).





## Hypergeometric pmf

Recall that the probability mass function (pmf) provides the probability of each element of the sample space. For a Hypergeometric random variable, there are a finite number of possible outcomes (successes):

$$S = \left\{ \max\left\{0, k-n\right\}, \ldots, \min\left\{k, m\right\} \right\}$$

Note that there are a few ways of parameterizing the Hypergeometric: by successes and total ($K$ and $N$) *or* by successes and failures ($K$ and $M$) *or* a combination of the two ($K$ and $N-K$).

But, they all have this form:

$$\mathbb{P}[\ X = x\ ] = \frac{\left(\text{the number of ways to succeed } x \text{ times}\right)\left(\text{the number of ways to fail}\right)}{\left(\text{the total number of ways in those } n \text{ trials}\right)}$$

## Hypergeometric pmf

From the previous equation, and from our knowledge of combinations, we can now see that the probability mass function can be written as

$$\mathbb{P}[\ X = x\ ] = \begin{cases} \dfrac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} & x \in S \\ 0 & \text{Otherwise} \end{cases}$$

Here,

$m$ = population successes
$n$ = population failures
$k$ = sample size

Not too confusing, right?





## Hypergeometric pmf

Or, we can use our knowledge of combinations and a different parameterization (Hawkes's) to see that the probability mass function can also be written as

$$\mathbb{P}[\ X = x\ ] = \begin{cases} \dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} & x \in S \\ 0 & \text{Otherwise} \end{cases}$$

Here,

$K$ = population successes
$N$ = population size
$n$ = sample size

Not too confusing, right?

## Hypergeometric Parameters

The expected value of a Hypergeometric is

$$\mathbb{E}\big[\,X\,\big] = k\,\frac{m}{n+m}$$

Similarly, the variance is

$$\mathbb{V}\big[\,X\,\big] = k\,\frac{m}{m+n}\,\frac{n}{m+n}\,\frac{m+n-k}{m+n-1}$$

Here,

$m$ = population successes
$n$ = population failures
$k$ = sample size





## Hypergeometric Parameters

The expected value of a Hypergeometric is

$$\mathbb{E}\big[\,X\,\big] = n\,\frac{K}{N}$$

Similarly, the variance is

$$\mathbb{V}\big[\,X\,\big] = n\,\frac{K}{N}\,\frac{N-K}{N}\,\frac{N-n}{N-1}$$

Here,

$K$ = population successes
$N$ = population size
$n$ = sample size

The **key** is to understand what the parameters *represent*. If you do this, then you will be able to move between the different parameterizations.

## Hypergeometric Parameters

There are two things to take away from these formulas when comparing the Hypergeometric to the Binomial distribution:

- The expected values are the same.
- The variance of the Hypergeometric is less than that of the Binomial.

You will want to keep this in mind as you do Lab Activity B.





## Hypergeometric Example 1: Spades

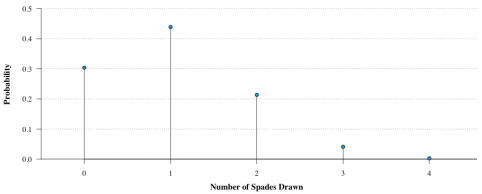
**Example**

Let $X$ be the number of spades drawn out of 4 draws from a deck of cards, without replacement. If the deck is fair, then it is clear that $X \sim \mathcal{H}\text{yper}(m = 13, n = 39, k = 4)$.

- What is the probability of getting one spade?
- What is the probability of getting at most three spades?
- What is the expected number of spades?

## Hypergeometric Example 1: Spades

The probability mass function (pmf) for the number of spades drawn:








## Hypergeometric Example 1: Spades

**Example**

Let $X$ be the number of spades drawn out of 4 draws from a deck of cards, without replacement. If the deck is fair, then it is clear that $X \sim \mathcal{H}\text{yper}(m = 13, n = 39, k = 4)$.

1. What is the probability of getting one spade?

We are asked to calculate $\mathbb{P}[\, X = 1 \,]$. This is a simple application of the complicated pmf:

$$\mathbb{P}[\, X = x \,] = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} = \frac{\binom{13}{1}\binom{39}{3}}{\binom{52}{4}}$$

$$= \frac{(13)(9139)}{270{,}725} = 0.438\,847$$

## Hypergeometric Example 1: Spades

In R, this is

```
dhyper(1, m=13, n=39, k=4)
```





## Hypergeometric Example 1: Spades

**Example**

Let $X$ be the number of spades drawn out of 4 draws from a deck of cards, without replacement. If the deck is fair, then it is clear that $X \sim \mathcal{H}\text{yper}(m = 13, n = 39, k = 4)$.

a. What is the probability of getting at most three spades?

We are asked to calculate $\mathbb{P}[\, X \leq 3 \,]$. This is a simple application of the complicated pmf:

$$\mathbb{P}[\, X \leq 3 \,] = \sum_{x=0}^{3} \frac{\binom{13}{x}\binom{39}{4-x}}{\binom{52}{4}} = \cdots$$

In R, this is

```
phyper(3, m=13, n=39, k=4)
```
$= 0.997\,358$

## Hypergeometric Example 1: Spades

**Example**

Let $X$ be the number of spades drawn out of 4 draws from a deck of cards, without replacement. If the deck is fair, then it is clear that $X \sim \mathcal{H}\text{yper}(m = 13, n = 39, k = 4)$.

1. What is the expected number of spades?

We are asked to calculate $\mathbb{E}\big[X\big]$. This is a simple application of the formula for the expected value:

$$\begin{aligned}\mathbb{E}\big[X\big] &= \text{sample size} \times \frac{\text{successes}}{\text{trials}} \\ &= 4\,\frac{13}{52} \\ &= 1\end{aligned}$$





## Hypergeometric Example 2: STAT 200

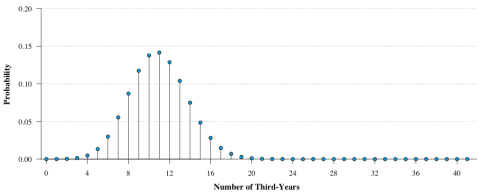
**Example**

I wonder if STAT 200 attracts Third-Year students at a greater rate than other courses. Let $X$ be the number of Third-Years in this STAT 200 course. From the Registrar's website, we know that the number of Third-Year students at Knox is $m = 356$ and non-Third-Years is $n = 978$. There are $k = 41$ students, of whom $x = 18$ are Third-Years.

Does this class support my claim?

## Hypergeometric Example 2: STAT 200

The pmf for the number of third-years in my STAT 200 section:








## Hypergeometric Example 2: STAT 200

**Solution**:
We are asked to calculate $\mathbb{P}[\, X \geq 18 \,]$.

Getting this in CDF format, this is $1 - \mathbb{P}[\, X \leq 17 \,]$.

In `R`, this is `1 - phyper(18, m=356, n=978, k=41)` $= 0.004\,739$.

**Interpretation**: Because this probability is so small, it appears as though Third-Years are over-represented in this course.

By the way, doing this calculation as if it were a Binomial random variable gives a p-value of $0.005\,372$.

## Hypergeometric Example 3: MATH 121

**Example**

I wonder if MATH 121 attracts Fourth-Year students at a lower rate than other courses. Let $X$ be the number of Fourth-Years in my MATH 121 course. From the Registrar's website, we know that the number of Fourth-Year students at Knox is $m = 278$ and non-Fourth-Years is $n = 1056$. There are $k = 30$ students in my MATH 121 course, of whom $x = 3$ are Fourth-Years.

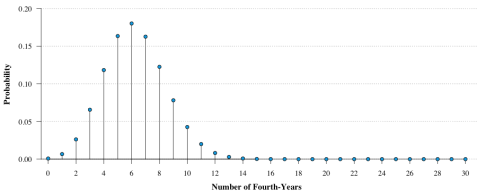
Does that class support my claim?





## Hypergeometric Example 3: MATH 121

The pmf for the number of fourth-years in my STAT 200 section:



## Hypergeometric Example 3: MATH 121

**Solution**:
We are asked to calculate $\mathbb{P}[\, X \leq 3 \,]$.

In R, this is `phyper(3, m=278, n=1056, k=30)` $= 0.099\,429$.

**Interpretation**: Because this probability is "not *that* small," there does not seem to be much evidence that MATH 121 attracts Fourth-Years at a lower rate than all other courses.

By the way, doing this calculation as if it were a Binomial random variable gives a p-value of $0.102\,043$. Not too different.





## Hypergeometric Example 4: The Unemployed

**Example**

I would like to form a committee of five of my Data Science alumni to review their Knox courses to determine which were most helpful in their current jobs. Unbeknownst to me, two of the 11 are currently unemployed. If I choose the committee members at random:

- What is the expected number of unemployed alumni in the committee?
- What is the probability that a majority is unemployed?
- What is the probability that all are employed?
- What is the middle 50% interval for the number of unemployed on the committee?

## Hypergeometric Example 4: The Unemployed

What are the values of $m$, $n$, and $k$?

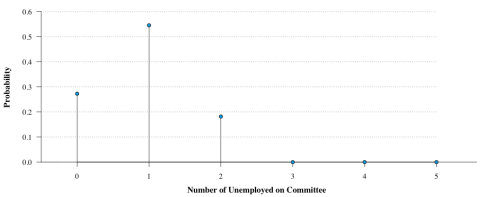
- Number of success in population: $m =?$
- Number of failures in population: $n =?$
- Number of people in the sample: $k =?$
- Number of success in population: $m = 2$
- Number of failures in population: $n = 9$
- Number of people in the sample: $k = 5$





## Hypergeometric Example 4: The Unemployed

The pmf for the number of unemployed alumni on the committee:



## Hypergeometric Example 4: The Unemployed

Here are the answers:

- What is the expected number of unemployed alumni in the committee?
  $= 4 \ \frac{2}{11} = 0.72727.$
- What is the probability that a majority is unemployed?
  `1 - phyper(2, m=2, n=9, k=5)` $= 0.000.$
- What is the probability that all are employed?
  `dhyper(0, m=2, n=9, k=5)` $= 0.2727.$
- What is the middle 50% interval for the number of unemployed on the committee?
  `qhyper( c(0.25,0.75), m=2, n=9, k=5)`
  We expect at least 50% of the possible committees to have between 0 and 1 unemployed members.





## Today's Objectives

Now that we have concluded this lecture, you should be able to

- determine what types of random variables follow a Hypergeometric distribution using its statistical definition
- calculate the usual two important population parameters:
  - expected value
  - variance
- identify the three parameters defining the Hypergeometric distribution
- calculate probabilities associated with a Hypergeometric random variable

## Today's R Functions

In this slide deck, we covered (or hinted) on the following four R functions related to the Hypergeometric distribution:

- `dhyper(x, m,n,k)` is the pmf, $\mathbb{P}[\ X = x\ ]$
- `phyper(x, m,n,k)` is the CDF, $\mathbb{P}[\ X \leq x\ ]$
- `qhyper(p, m,n,k)` calculates $x$, such that $\mathbb{P}[\ X \leq x\ ] = p$ (the **q**uantile function)
- `rhyper(n, m,n,k)` generates a random sample from this distribution (the **r**andom function)

**Please** do not forget to access the `allProbabilities` document that provides all of the important probability functions in R.





## Supplemental Activities

The following are supplements for the topics covered today.

- SCA 5a is for some discrete distributions

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCA, **Laboratory Activity B** is helpful for learning how to handle discrete distributions (including the Hypergeometric distribution). The lab actually shows the connection between sampling and discrete distributions. It uses three named distributions.
`https://www.kvasaheim.com/courses/stat200/labs/`

## Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding discrete distributions:

- Hawkes Learning: Section 5.4
- Intro to Modern Statistics: None
- R for Starters: Appendix A.6

- Wikipedia: Hypergeometric Distribution