Module C: Understanding the Data-Generating Process

Slide Deck C5:

Geometric Distributions

The section in which we learn about another named discrete distributions in which the random variable is the number of trials. A random variable that follows a Geometric distribution will model the number of failures until the first success.

Today's Objectives

By the end of this slidedeck, you should

1. determine what types of random variables follow a Geometric distribution using its statistical definition
2. calculate the expected value and variance
3. identify the one parameter defining the Geometric distribution
4. calculate probabilities associated with a Geometric random variable
5. distinguish between a Geometric and a Binomial random variable

Definition of a Geometric Experiment

Definition

The **Geometric distribution** describes the probability of obtaining *one* success after X failures.

Examples

- number of times you need to draw a card (with replacement) before you get your first heart
- number of people you need to ask before you meet your first Sophomore (with replacement)
- number of experiments run until one fails (where each has a 5% chance of failing)

The difference between a Bernoulli and a Geometric distribution is that the Bernoulli specifies the number of trials, while the Geometric treats that as the random variable.

Geometric pmf

Recall that the probability mass function (pmf) provides the probability of each element of the sample space. For a Geometric random variable, there are an infinite number of possible outcomes (failures until first success):

$$S = \{0, 1, 2, \infty\}$$

Note that the pmf of a Geometric random variable is easily devised. Recall for the Binomial distribution:

$$\mathbb{P}[\ X = x\ ] = \underbrace{\binom{n}{x}}_{\text{combinatons}} \underbrace{p^x}_{\text{successes}} \underbrace{(1-p)^{n-x}}_{\text{failures}}$$

Geometric pmf

For the Geometric, we know there is just one success — but the last is fixed at the end. Thus, the pmf is

$$\mathbb{P}[\ X = x\ ] = \underbrace{\binom{n}{x}}_{\text{combinatons}} \underbrace{p^x}_{\text{successes}} \underbrace{(1-p)^{n-x}}_{\text{failures}} \qquad \text{Binomial}$$

$$\mathbb{P}[\ X = x\ ] = \underbrace{1}_{\text{combinatons}} \underbrace{p^1}_{\text{successes}} \underbrace{(1-p)^{x}}_{\text{failures}} \qquad \text{Geometric}$$

In the Geometric, x is the **number of failures** until the first success.

Example 1: The Ace of Spades

Example

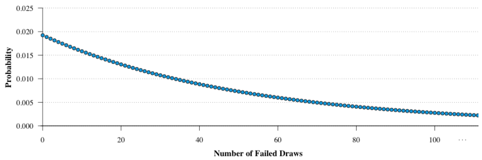
Let X be the number of times I draw a card (with replacement) from a standard deck of cards until I get the Ace of Spades. If the deck is fair, then it is clear that

$$X \sim \mathcal{G}eom\,(p = 1/52)$$

1. What is the probability of getting the first Ace of Spades on the first draw?
2. What is the probability of getting the first Ace of Spades on the 10th draw?
3. What is the probability of getting the first Ace of Spades on the 100th draw?

Example 1: The Ace of Spades

The probability mass function of the $\mathcal{G}eom\,(p = 1/52)$ distribution:



Example 1: The Ace of Spades

Example

Let X be the number of times I draw a card (with replacement) from a standard deck of cards before I get the Ace of Spades. If the deck is fair, then it is clear that

$$X \sim \mathcal{G}eom\,(p = 1/52)$$

1. What is the probability of getting the Ace of Spades on the first draw?

Remember that the Geometric models the number of *failures* until the first success. So, since we are asked to calculate a probability about getting the Ace of Spades on the first draw, we are asked to calculate the probability of having 0 failures.

Example 1: The Ace of Spades

That is, we are asked to calculate $\mathbb{P}\left[\ X = 0 \ \middle|\ p = 1/52\ \right]$. This is a simple application of the pmf.

$$\frac{1}{52} \times \left(\frac{1}{52}\right)^{0}$$

But, because we realize the computer can perform the calculation much more easily, we use this in R:

```
dgeom(0, 1/52)
```


Start of Lecture Material
Geometric Distributions
Some Examples
End of Lecture Material
Example 1: The Ace of Spades
Example 2: Those Sophomores
Example 3: Experimenting

Example 1: The Ace of Spades

Example

Let X be the number of times I draw a card (with replacement) from a standard deck of cards before I get the Ace of Spades. If the deck is fair, then it is clear that

$$X \sim \mathcal{G}eom\left(p = 1/52\right)$$

2. What is the probability of getting the first Ace of Spades on the 10th draw?

We are asked to calculate $\mathbb{P}\left[\ X = 9 \ \middle|\ p = 1/52\ \right]$. In R, this is

`dgeom(9, 1/52)` $= 0.016\,147\,230$

Example 1: The Ace of Spades

Example

Let X be the number of times I draw a card (with replacement) from a standard deck of cards before I get the Ace of Spades. If the deck is fair, then it is clear that

$$X \sim \mathcal{G}eom\,(p = 1/52)$$

3. What is the probability of getting the first Ace of Spades on the 100th draw?

We are asked to calculate $\mathbb{P}\left[\ X = 99\ \middle|\ p = 1/52\ \right]$. In R, this is
`dgeom(99, 1/52)` $= 0.002\,812\,633$

Example 2: Those Sophomores

Example

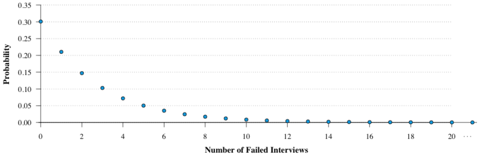
Ultimately, I would like to estimate the number of Sophomores at Knox College. However, for this example, let's look at how variable the samples are. We know that the number of Sophomores is 365 out of a total of 1213 students.

I randomly wander campus and ask students their year in school. I stop when I meet my first Sophomore.

1. What is the probability that the first student I ask is a Sophomore?
2. What is the probability that the 10th student I ask is my first Sophomore?
3. What is the probability that I will ask my first Sophomore *by the time* I ask my 10th student?

Example 2: Those Sophomores

The probability mass function of the $\mathcal{G}eom\,(p = 365/1213)$ distribution:



Example 2: Those Sophomores

Example

Let S be the number of times I interview a student who is not a Sophomore. If I am random in my selection of students, then it is clear that

$$S \sim \mathcal{G}eom\,(p = 365/1213)$$

1. What is the probability that the first student I ask is a Sophomore?

Remember that the Geometric models the number of *failures* until the first success. So, since we are asked to calculate a probability about getting the Sophomore on the first draw, we are asked to calculate the probability of having 0 failures.

Example 2: Those Sophomores

That is, we are asked to calculate $\mathbb{P}\left[\ S = 0 \ \middle|\ p = 365/1213\ \right]$. This is a simple application of the pmf (and of logic). But, because we realize the computer can perform the calculation much more easily, we use this in R:

`dgeom(0, 365/1213)` $= 0.300\,907$

Example 2: Those Sophomores

Example

Let S be the number of times I interview a student who is not a Sophomore. If I am random in my selection of students, then it is clear that

$$S \sim \mathcal{G}eom\,(p = 365/1213)$$

2. What is the probability that the first Sophomore is the tenth student?

That is, we are asked to calculate $\mathbb{P}\left[\ S = 9 \ \middle|\ p = 365/1213\ \right]$. This is a simple application of the pmf. But, because we realize the computer can perform the calculation much more easily, we use this in R:

`dgeom(9, 365/1213)` $= 0.012\,009$

Example 2: Those Sophomores

Example

Let S be the number of times I interview a student who is not a Sophomore. If I am random in my selection of students, then it is clear that

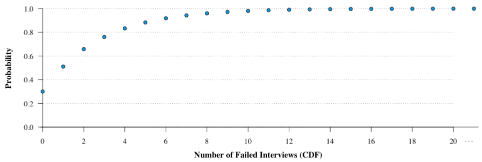
$$S \sim \mathcal{G}eom\,(p = 365/1213)$$

3. What is the probability that I will ask my first Sophomore by the time I ask my 10th student?

That is, we are asked to calculate $\mathbb{P}\left[\ S \leq 9 \ \middle|\ p = 365/1213\ \right]$. This is a simple application of the CDF.

Example 2: Those Sophomores

Since we are asking about "$\leq$", we need to use the cumulative distribution function (CDF). This is a graphc of the CDF of the $\mathcal{G}eom\,(p = 365/1213)$ distribution:



Example 2: Those Sophomores

That is, we are asked to calculate $\mathbb{P}\left[\ S \leq 9 \ \middle|\ p = 365/1213\ \right]$. This is a simple application of the CDF:

`pgeom(9, 365/1213)` $= 0.972\,116$

Example 3: Experimenting

Example

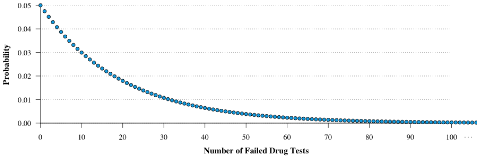
For reasons that will become clear in the not-so-distant future, the probability that an experiment will "detect a significant difference" is 5%. Let us explore the consequences of this.

Using pharmaceuticals, I would like to improve the life of people with lower back problems. To do this, I perform experiments on each of my drugs. The probability that any one of my statistical analyses (drug trials) has of concluding the drug seems to work is 5%.

1. What is the probability that I conclude the first drug works?
2. What is the probability that I conclude the first 99 do not work, but that 100th works?
3. What is the probability that I conclude a drug works by the 20th drug?

Example 2: Experimenting

The probability mass function of the $\mathcal{G}eom\,(p = 0.05)$ distribution:



Example 3: Experimenting

Example

Using pharmaceuticals, I would like to improve the life of people with lower back problems. To do this, I perform experiments on each of my drugs. The probability that any one of my statistical analyses (drug trials) has of concluding the drug seems to work is 5%.

1. What is the probability that I conclude the first drug works?

That is, we are asked to calculate $\mathbb{P}\left[\ D = 0\ \middle|\ p = 0.05\ \right]$. This is a simple application of the pmf (and of logic):

`dgeom(0, 0.05)` $= 0.05$

Example 3: Experimenting

Example

Using pharmaceuticals, I would like to improve the life of people with lower back problems. To do this, I perform experiments on each of my drugs. The probability that any one of my statistical analyses (drug trials) has of concluding the drug seems to work is 5%.

2. What is the probability that I conclude the first 99 do not work, but that 100th works?

That is, we are asked to calculate $\mathbb{P}\left[\ D = 99 \ \middle|\ p = 0.05\ \right]$. This is a simple application of the pmf:

`dgeom(99, 0.05)` $= 0.000\,311\,607$

Example 3: Experimenting

Example

Using pharmaceuticals, I would like to improve the life of people with lower back problems. To do this, I perform experiments on each of my drugs. The probability that any one of my statistical analyses (drug trials) has of concluding the drug seems to work is 5%.

3. What is the probability that I conclude some drug works by the 20th drug?

That is, we are asked to calculate $\mathbb{P}\left[\ D \leq 20 \ \middle|\ p = 0.05\ \right]$. This is a simple application of the CDF:

`pgeom(20, 0.05)` $= 0.659\,438$

Today's Objectives

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

1. determine what types of random variables follow a Geometric distribution using its statistical definition
2. calculate the expected value and variance
3. identify the one parameter defining the Geometric distribution
4. calculate probabilities associated with a Geometric random variable
5. distinguish between a Geometric and a Binomial random variable

Useful R Functions

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

- `dgeom(x, p)` is the pmf, p for: $\mathbb{P}[\ X = x\ ] = p$
- `pgeom(x, p)` is the CDF, p for: $\mathbb{P}[\ X \leq x\ ] = p$
- `qgeom(q, p)` is the quantile function, x for: $\mathbb{P}[\ X \leq x\ ] = p$
- `rgeom(n, p)` generates n random values from: $\mathcal{Geom}(\texttt{prob}{=}p)$

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

Supplemental Activities

The following are supplements for the topics covered today.

- SCA 5 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 Geometric 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 the Geometric discrete distribution:

- Hawkes Learning: None
- Intro to Modern Statistics: None
- R for Starters: Appendix A.4
- Wikipedia: Geometric Distribution