This class is being conducted over Zoom. As the instructor, I will be recording this session. I have disabled the recording feature for others so that no one else will be able to record this session. I will be posting this session to the course's website.

If you have privacy concerns and do not wish to appear in the recording, you may turn video off (click "stop video") so that Zoom does not record you.

The chat box is always open for discussion and questions to the entire class. You may also send messages privately to the instructor or the TAs. Please note that Zoom saves all chat transcripts.

I create a live transcription of each session using Otter.ai. This means that Otter.ai will transcribe anything spoken over the Zoom audio. The transcript will be posted with the session video on the course website.

Discrete Probability Distributions

Stats 7

Mary Ryan

Aug. 20, 2020

Course website:

https://canvas.eee.uci.edu/courses/28451

Slides can be found at:

https://maryryan.github.io/stats7-SS2-2020-slides/stats7-SS2-2020-discreteDist/stats7-SS2-2020-discreteDist

Learning Objectives

By the end of today's lecture, you should be able to:

- identify whether a probability distribution is discrete
- differentiate between 4 major types of discrete probability distributions: Uniform, Binomial, Geometric, and Poisson
- calculate probabilities, expected values, and variances for the 4 major types of discrete probability distributions
- understand how to model real-world events with discrete probability distributions

Probability Distributions

- Previously, probability for events have been given to us
- What if we want to know the probability of an event, but have no data?
- Might be useful to apply a **probability model** or **probability distribution**
    - Existing framework with **known properties**, given that certain **conditions** apply
    - Each distribution has formulae for calculating probability
    - Different types of models for **discrete** and **continuous** variables
    - We call these variables, **random variables**
        - Today, we'll focus on discrete random variables

Discrete Random Variables

- Like discrete data types, discrete random variables are variables that take on numerical values in **jumps**
- However, we believe the values these variables take on is **random**
    - We try to identify the random process that generates these values by looking at **probability distributions**

Probability Distributions

- How to pick a probability distribution?
    - Compare the distribution's conditions and assumptions to your scenario and see if they apply

Probability Distributions

- How to pick a probability distribution?
  - Compare the distribution's conditions and assumptions to your scenario and see if they apply
- Think of a scenario where you would like to find some probabilities like a unidentifiable article of clothing
  - You have no knowledge of what the clothing item is actually *meant* to be

Probability Distributions

- Think of a scenario where you would like to find some probabilities like a unidentifiable article of clothing
    - You examine the item of clothing and compare it to assumptions and conditions you have about clothing items you can identify

I know pants have
2 holes for legs,
so it can't be that.

Probability Distributions

- Think of a scenario where you would like to find some probabilities like a unidentifiable article of clothing
    - Once you make your comparisons, you identify the unknown item as an article of clothing that meets the most conditions, because that's the best you can do

Probability Distributions

- Think of a scenario where you would like to find some probabilities like a unidentifiable article of clothing
    - Once you make your comparisons, you identify the unknown item as an article of clothing that meets the most conditions, because that's the best you can do

- We do the same with probability distributions:
    - We can **rarely** know the function that **truly** generates probabilities for a scenario
    - To give it our **best guess**, we see which probability distribution our scenario meets most of the conditions for, and use those functions and known properties

Discrete Uniform Distribution

- The most simple probability distribution we might think of is one where all events have an **equal** probability of happening

Discrete Uniform Distribution

- The most simple probability distribution we might think of is one where all events have an **equal** probability of happening

- Applies to **discrete** random variables
    - Variables can take on values that exist between some number *a* and another number *b*
    - Each value is **equally likely** to happen
    - Defined by the minimum value the variable can take on (*a*) and the maximum value it can take on (*b*)

- Some properties:
    - $P(X = x) = \frac{1}{b-a+1} = \frac{1}{\text{\# of total values random variable X can take on}}$ (known as the **probability mass function**, or pmf)
    - $E(X) = \frac{a+b}{2}$
    - $Var(X) = \frac{(b-a+1)^2-1}{12}$

- An evenly weighted die

Dice Activity (10 minutes)

- Split up into breakout room groups of 3 (1 spokesperson, 1 drawer, 1 recorder)
- Each group is assigned a die: d4 (groups 1, 6, 11, 16), d6 (groups 2, 7, 12, 17), d8 (groups 3, 8, 13, 18), d10 (groups 4, 9, 14, 19), d20 (groups 5, 10, 15, 20)
- Go to Google and search "roll _", with your type of die in the blank
  - What value did you get? Record it
- Take turns "rolling" the die and recording the values you get. Do this until you have 100 observations
  - You can roll dice in bulk. To roll 10 d10 dice at once, search "roll 10 d10". It caps you at rolling 90 dice at one time
- What are the observed probabilities for each die face value?
- Create a histogram of your observations. Describe its shape
- Explain why a Uniform distribution might be a good model for scenarios involving your die

A Game of Dice

- Say we are playing a game with a regular six-sided die
  - If we roll a 4 or higher, we win
  - If we roll a 3 or lower, we lose

A Game of Dice

- Say we are playing a game with a regular six-sided die
  - If we roll a 4 or higher, we win
  - If we roll a 3 or lower, we lose
- What is the probability of winning? What is the probability of losing?

A Game of Dice

- Say we are playing a game with a regular six-sided die
  - If we roll a 4 or higher, we win
  - If we roll a 3 or lower, we lose
- What is the probability of winning? What is the probability of losing?
- We know a die can be described as a discrete Uniform distribution. Can we describe this game as a Uniform?

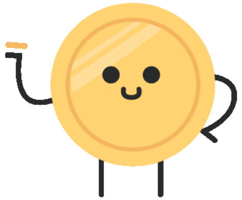
Getting More Complicated

- The last example shows that sometimes we aren't really interested in the probability of *any* event happening, but just a particular (set of) event(s)
- We can split up all the events into two mutually exclusive groups: "events we're interested in" and "events we're not interested in"
    - We can still describe this as a Uniform distribution if the event(s) we're interested in have the same probability those event(s) not happening

Getting More Complicated

- The last example shows that sometimes we aren't really interested in the probability of *any* event happening, but just a particular (set of) event(s)
- We can split up all the events into two mutually exclusive groups: "events we're interested in" and "events we're not interested in"
    - We can still describe this as a Uniform distribution if the event(s) we're interested in have the same probability those event(s) not happening
- What if we don't necessarily think that all events have the same probability of happening?

Bernoulli Distribution



- Applies to **discrete** random variables
  - Variable can be described as whether an event happens (**"success"**) in a trial
  - Each trial has a known probability of "success", *p*
  - Can take on values that exist between 0 (no success observed) and 1 (a success observed)

- Some properties:
  - $P(X = x) = p^x(1 - p)^{1-x}$
  - $E(X) = p$
  - $Var(X) = p(1 - p)$

- Number of heads you get in a single coin flip
  - Each head is a "success"
  - Probability of success = 0.5 for each trial

Back to our dice game

- Say we are playing another game with a regular six-sided die
    - If we roll a 5 or higher, we win
    - If we roll a 4 or lower, we lose

Back to our dice game

- Say we are playing another game with a regular six-sided die
  - If we roll a 5 or higher, we win
  - If we roll a 4 or lower, we lose
- What is *p*?
- What is the probability of winning the game? Of losing?

Back to our dice game

- Say we are playing another game with a regular six-sided die
  - If we roll a 5 or higher, we win
  - If we roll a 4 or lower, we lose
- What is p?
- What is the probability of winning the game? Of losing?
- What if we wanted to play this game several times?

Binomial Distribution

- Applies to **discrete** random variables
  - Variable can be described as the number of "successes" in *n* **independent trials**
  - Each trial has a known probability of "success", *p*
  - Can take on values that exist between 0 (no successes observed) and *n* (all trials were a success)
- It's like we're doing *n* Bernoulli trials

- Some properties:
  - $P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}$
  - $E(X) = np$
  - $Var(X) = np(1-p)$

- Number of heads you get in 10 coin flips
  - n=10 independent trials
  - Each head is a "success"
  - Probability of success = 0.5 for each trial

Binomial Distribution

- What's up with $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$?

Binomial Distribution

- What's up with $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$?
- When performing the same trial over and over again, there are **many different ways** to get x successes
  - All the successes could be seen at the beginning of the trials, or at the end, or in the middle
- Can think of $P(X=x)$ as:

$$P(Event) = (\#\text{ scenarios event can occur})P(\text{Single scenario})$$

Binomial Distribution

- What's up with $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$?
- When performing the same trial over and over again, there are **many different ways** to get *x* successes
  - All the successes could be seen at the beginning of the trials, or at the end, or in the middle
- Can think of $P(X = x)$ as:

$$P(Event) = (\#\text{ scenarios event can occur})P(\text{Single scenario})$$

- The **choose function**, $\binom{n}{x}$, gives us the total number of ways of selecting *x* distinct combinations of *n* trials, irrespective of order
- Since all the trials are **independent**, we can use the independence property $P(\text{A and B}) = P(A)P(B)$:
  - $p^x(1-p)^{n-x}$, tells us the probability that *x* successes occur and (*n*-*x*) failures occur

Example 1: Two-Factor Authentication

A June 2019 survey by Pew Research Center found that only 28% of U.S. adults were able to correctly identify an example of two-factor authentication.

To see if this holds true at UCI, we provide the same question to 20 UCI students.

- To think of this scenario in terms of the Binomial distribution, what are the trials we are performing? Are they independent?

- What can we think of as a "success"? What is the probability of success?

Example 1: Two-Factor Authentication

A June 2019 survey by Pew Research Center found that only 28% of U.S. adults were able to correctly identify an example of two-factor authentication.

To see if this holds true at UCI, we provide the same question to 20 UCI students.

- If the Pew Research Center proportion is true for the general population, what can we expect the probability to be that:
    - 13 of our UCI students are able to correctly identify two-factor authentication?
    - 5 of our UCI students are able to correctly identify two-factor authentication?

Calculator: binompdf()

To get to the calculator function on a TI-84:

- 2nd DISTR > A: binompdf(

To calculate P(X=a): binompdf(n,p,a)

Example 1: Two-Factor Authentication

A June 2019 survey by Pew Research Center found that only 28% of U.S. adults were able to correctly identify an example of two-factor authentication.

To see if this holds true at UCI, we provide the same question to 20 UCI students.

- If the Pew Research Center proportion is true for the general population, what can we expect the probability to be that:
    - Fewer than 3 UCI students are able to correctly identify two-factor authentication?
    - At least 18 UCI students are able to correctly identify two-factor authentication?

Calculator: binomcdf()

To get to the calculator function on a TI-84:

- 2nd DISTR > B: binomcdf(

To calculate P(X $\leq$ a):

- binomcdf(n,p,a)

Example 2: Data Breach

A June 2019 survey by Pew Research Center found that 52% of U.S. adults recently decided not to use a product or service because they were worried about how much personal information would be collected about them.

- Let's say that we enroll 50 Orange County residents in our study. Based on our prior knowledge from Pew Research, how many might we expect to say decided not to use a product or service because because of personal information collection?

Example 2: Data Breach

A June 2019 survey by Pew Research Center found that 52% of U.S. adults recently decided not to use a product or service because they were worried about how much personal information would be collected about them.

- From our 50 study enrollees, what is the probability that:
    - 30 decided not to use a product or service because because of personal information collection?
    - More than 20, but fewer than 27 decided not to use a product or service because because of personal information collection?

Example 3: Household Languages

According to the 2018 American Community Survey from the U.S. Census Bureau, 21.9% of U.S. households speak a language other than English at home.

- If 3,285 people in the survey said their household speaks a language other than English at home, how many people in total were surveyed?

- Do we have independent trials here? What issues may put this independence into jeopardy?

Dice Game, Part 3

- Say we are playing another game with a regular six-sided die
    - If we roll a 5 or higher, we win
    - If we roll a 4 or lower, we lose

Dice Game, Part 3

- Say we are playing another game with a regular six-sided die
  - If we roll a 5 or higher, we win
  - If we roll a 4 or lower, we lose
- If we play this game 5 times, what is the probability we win at least once?
- What if we say we're going to keep playing until we win?

Geometric Distribution

- Applies to **discrete** random variables
  - Variable can be described as the number of trials it takes to observe a success
  - Known probability of success, *p*, in each trial is the same
  - Can take on values that exist between 1 (get a success of the first trial) and ∞

- Some properties:
  - $P(X = x) = p(1-p)^{n-1}$
  - $E(X) = \frac{1}{p}$
  - $Var(X) = \frac{1-p}{p^2}$

- Number of coin flips it takes you to get 1 tail
  - Each tail is a "success"
  - Probability of success = 0.5 for each trial

A Warning

- Be careful when you look up the Geometric distribution online
    - You can also express this distribution in terms of the number of failures you need in order to get a success
    - This changes the formula for $P(X = x)$, the expected value, and the variance
- If you see a problem being worked out in a way that doesn't match up with these notes, it's likely that they're using the other parameterization
    - Both ways are correct, just be careful whether you're looking at number of total trials or number of failed trials to make sure you're not mixing up methods

Example 1: UCI First Gen

According to U.S. News & World Report, 44% of UCI undergraduates in 2017 were first generation college students. We are running a study looking at the experiences of first generation college students at UCI, and need to find eligable students to enroll.

- If we randomly choose UCI undergraduates to contact from the university masterlist, what is the probability that any given student will be first generation?

- If we contact 10 random students, what is the probability that we would contact at least 1 first generation student?

Example 1: UCI First Gen

According to U.S. News & World Report, 44% of UCI undergraduates in 2017 were first generation college students. We are running a study looking at the experiences of first generation college students at UCI, and need to find eligable students to enroll.

- What is the probability that the first first-gen student we get in contact with is the 5th person we've contacted today?

- What is the probability that we will contact a first generation student within the first 3 students we contact?

Calculator: geometpdf() & geometcdf()

To get to the calculator function on a TI-84:

- 2nd DISTR > E: geometpdf(
- 2nd DISTR > F: geometcdf()

To calculate P(X=a):

- geometpdf(p,a)

To calculate P(X $\leq$ a):

- geometcdf(p,a)

Example 2: Minecraft Creepers

In the game Minecraft, creepers are green monsters that will explode when close to a player. When you kill this monster, it will sometimes drop items like gunpowder or music discs. There is a 67% chance that the creeper will drop at least one unit of gunpowder when you kill it.

- How many creepers can we expect to kill before we get our first gun powder drop?

- What is the probability that we will get our first gun powder drop when we kill the 4th creeper?

Example 2: Minecraft Creepers

In the game Minecraft, creepers are green monsters that will explode when close to a player. When you kill this monster, it will sometimes drop items like gunpowder or music discs. There is a 67% chance that the creeper will drop at least one unit of gunpowder when you kill it.

- What is the probability that we will kill no more than 6 creepers to get our first gunpowder drop?

- What is the probability that we will need to kill at least 2 creepers and at most 4 creepers to get our first gun powder drop?

Poisson Distribution

- Applies to **discrete** random variables
  - Variable can be described the number of events that occur in a **defined period of time**
  - The rate at which events happen is known as $\lambda = \frac{\text{\# events}}{\text{amount of time}}$
  - Can take on values that exist between 0 and ∞

- Some properties:
  - $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$
  - $E(X) = \lambda$
  - $Var(X) = \lambda$

- The number of cars that pass through an intersection in 1 hour
  - $\lambda = \frac{\text{\# cars}}{\text{1 hour}}$

Example 1: Nuclear Plant

Say a particular nuclear plant releases a detectable amount of radioactive gases twice a month, on average.

- What is λ? What are our time units?

- How many times can we expect the plant to release a detectable amount of gas in a typical 6-month span?

Example 1: Nuclear Plant

Say a particular nuclear plant releases a detectable amount of radioactive gases twice a month, on average.

- What is the probability that the plant will release a detectable amount of gas 4 times in a typical month?

- What is the probability that the plant will release a detectable amount of gas fewer than 3 times in a typical month?

Calculator: poissonpdf() & poissoncdf()

To get to the calculator function on a TI-84:

- 2nd DISTR > C: poissonpdf(
- 2nd DISTR > D: poissoncdf()

To calculate P(X=a):

- poissonpdf(lambda,a)

To calculate P(X $\leq$ a):

- poissoncdf(lambda,a)

Example 1: Nuclear Plant

Say a particular nuclear plant releases a detectable amount of radioactive gases twice a month, on average.

- What is the probability that the plant will release a detectable amount of gas more than once but no more than 5 times in a typical month?