CMPT 155: Computer Applications for Life Sciences

Lecture 10: Discrete Probability

Ivan E. Perez

March 29, 2022

Presentation Outline

Homework & Administrative Schedule

Older Action Items

- Homework 5 Due: Friday, March 25th at 6 pm
- Second Midterm:
    - Practice test: Wednesday, March 23rd
    - Second Midterm Review: Friday March 25th
    - Second Midterm Exam: Tuesday, March 29th

Newer Action Items

- Last Day to Drop Classes: Wednesday April 20th
- Last Day of Classes: Friday May 6th
- Final Exams:
    - Section 01 (8am) Final Exam: May 9th 11am - 1pm
    - Section 02 (9am) Final Exam: May 10th 11 am - 1pm

See Registrar's Final Exam schedule for more information.

Fundamentals of Probability

Discrete probability is about measuring the subset of outcomes that satisfy the restriction amongst a broader set of outcomes.

Some definitions:

- Event:
  - An observation (e.g., coin flip, card draw, dice roll)
- Outcome:
  - Result of an event (e.g., Heads/Tails(H/T), King of Hearts (K♡), [1,6])
- Probability:
  - "Chance" of the occurrence of an event. (BOUNDED BETWEEN 0 and 1)
- Sample Space
  - Set of all possible outcomes of an experiment. (e.g., H/T, [1,6])

Some more Definitions

- Independent Events:
  - An event is called independent IF the outcome of one event does not influence the outcome of the other
- Mutually Exclusive Events/Outcomes:
  - Outcomes of Events that can not result in observing multiple outcomes at the same time.

Priori Probability

The PROBABILITY of an Event(s) occuring is the COUNT of Event(s) of Interest divided by the Sample Space (i.e., the set of all possible outcomes)

Using notation using a coinflip:

$$p(H) = \frac{\mathsf{H}}{\mathsf{H} + \mathsf{T}} = \frac{1}{2}$$

We can also compute the probability of rolling a $\boxed{4}$ on a 6-sided die:

$$p\left(\boxed{4}\right) = \frac{\boxed{4}}{\boxed{1}, \boxed{2}, \boxed{3}, \boxed{4}, \boxed{5}, \boxed{6}} = \frac{1}{6}$$

- Priori Probabilities can be computed *without* experimentation.
- Games (rolling dice)
- Examples: Lottery Odds, Trading card rarity odds, poker/blackjack odds.

Empirical Probability

The PROBABILITY of an Event(s) occurring is the COUNT of Event(s) of Interests divided by the total COUNT of Events observed Using notation watching 3 people enter the subway out of 10 people.

$$p(\text{`Enter the Subway'}) = \frac{3}{10}$$

- Computed from observations or event occurrences.
- Examples: Surveys, Store visits, Empirical Experiments

Example 1: SimpleProbability.xlsx

1. Download *'SimpleProbability.xlsx'*
2. Compute the *proportions* for each set of families with 0,1,...,7 children.
3. Answer the following questions:
   - Compute the probability that a family selected in this study will have **no** children?
   - Compute the probability that a family selected in this study will have **at least 4** children?

Example 1: Solution Picture

	A	B	C	D	E	F
1	Children	Families	Proportion		Families with	Proportion
2	0	256	0.10015649		No Children	0.1001565
3	1	682	0.26682316		At Least 4 children	0.1138498
4	2	796	0.3114241			
5	3	531	0.20774648			
6	4	221	0.08646322			
7	5	59	0.02308294			
8	6	10	0.00391236			
9	7	1	0.00039124			

What have we noticed about Probability so far?

1. Probabilities of mutually exclusive events can be added together.
2. The sum of the probabilities of all possible mutually exclusive events will always be $\mathbf{1}$.

Exercise 1: Probability of Rolling Fair Dice

On a new spreadsheet layout all the possible outcomes of two rolls of a fair die.

Compute the following probabilities:

- outcome of the dice roll will add up to 7.
- outcome of the dice roll will add up to 8 or more.

Exercise 1: Solution

1. Begin by filling Column A with the range of values that two dice can take
2. In column B list the ways each value can be observed by the dice, (e.g., 2 = snake eyes, or (1,1))
3. In Column C count the different ways each value can be observed by the dice.
4. In Column D compute the probability of each outcome by dividing count of the ways each outcome can be observed by the total count of all the different outcomes.

To compute probability ranges:

- outcome of the dice roll will add up to 7.
  - See Cell D7
- outcome of the dice roll will add up to 8 or more.
  - Add up the computed probabilities in cells D8:D12

Exercise 1: Solution (continued)

	A	B	C	D	E	F	G
1	Outcomes	Ways	Count of Ways	Proportion		Probability outcome =	
2	2	(1,1)	1	0.028		7	0.167
3	3	(1,2), (2,1)	2	0.056		8 or more	0.417
4	4	(2,2), (1,3), (3,1)	3	0.083			
5	5	(2,3), (3,2), (4,1), (1,4)	4	0.111			
6	6	(3,3), (4,2), (2,4), (5,1), (1,5)	5	0.139			
7	7	(1,6), (6,1), (5,2), (2,5), (4,3), (3,4)	6	0.167			
8	8	(4,4), (5,3), (3,5), (6,2), (2,6)	5	0.139			
9	9	(5,4), (4,5), (6,3), (3,6)	4	0.111			
10	10	(5,5), (6,4), (4,6)	3	0.083			
11	11	(5,6), (6,5)	2	0.056			
12	12	(6,6)	1	0.028			

Exercise 2a: Binomial Modelling of Emergency Room Arrivals

1. Layout a spreadsheet that has all the possible outcomes of a visit by three people to an emergency room.
2. Compute the following probabilities out of three arrivals:
   - **No** emergencies.
   - **One** emergency.
   - **Two** emergencies.
   - **Three** emergencies.

Exercise 2a: Solution

1. Create a table that will hold all the possible order of people arriving to the emergency room
   - Add the headers for arrivals 1,2, and 3 in cells A1:C1.
   - In Cells A2:C9 write out all the possible sequences of arrivals.
2. Create a table that counts the *ways* each emergency count is observed.
   - In Cells D1:F1, add the headers for:
     1. The Count of Emergencies for each case of arrivals,
     2. Count of ways to get to each case of arrivals,
     3. The Proportion/Probability of each case of arrivals.
3. In Cells D2:D5 list the emergencies count for each case.
4. In Cells E2:E5 list the count of ways to get to each emergency case.
5. In Cells F2:F5 compute the proportion of ways that get to the case against out of the total ways.

Exercise 2a: Solution (continued)

	A	B	C	D	E	F
1	Arrival 1	Arrival 2	Arrival 3	Emergencies	Count Ways	Proportion
2	O	O	O	0	1	0.14285714
3	E	O	O	1	3	0.42857143
4	O	E	O	2	2	0.28571429
5	O	O	E	3	1	0.14285714
6	E	E	O			
7	O	E	E			
8	E	O	E			
9	E	E	E			

In the Next Exercise (2b), let's try using *Emipirical* Probabilities based on data observed by a local hospital.

Exercise/Example 2b: Emergencies *Empirical*

1. Download *Emergencies.xlsx*
2. In Sheet2, Cells B12 and B13 compute the *Empirical* probability of an Emergency and an Other (non-emergency) using the data.
3. Using these computed probabilities fill in the probabilities of each independent event in Cells D3:F10.
4. In Cells H3:H10 fill in the count of the number of emergencies in each of the listed sequence of outcomes.
5. Lets stop here and we will learn about the 'Law of Independence' to compute the Final probabilities in cell G3:G10.

The Law of Independence

Assuming that the outcome of one event has no influence on successive outcomes then the joint probability (i.e., the probability of the previous outcome and the next outcome) is the product of the probabilities of each individual outcome.

Using coinflipping as an example If the probability of

1. getting a heads is $p(H) = 1/2$
2. getting a tails is $p(T) = 1/2$, and
3. the coinflips are assumed to be independent.

The probability of flipping a coin twice and getting a heads followed by a tails is:

$$p(H, T) = p(H) * p(T) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

Exercise/Example 2b: Emergencies *Empirical*

Continuing from Step 5

1. Use the ‘Law of Independence’ and the table you created in D3:F10 to compute the Final Probabilities in cells G3:G10.
2. Use our understanding of *mutually exclusive* events/outcomes to compute the probabilities of each case in cells K3:K6.
3. Save this sheet for when we use BINOMDIST() later on.

Binomial Probability

We define a *binomial distrubtion* to be an experiment in which:

- there are a fixed number of n-independent trials.
- each trial has only 2 possible outcomes, Success or Failure
  - We are interested in counting the number of k-successes.
- the probability of a success, p, for a single trial remains fixed throughout the experiement.

The *Expected Value* of Successes, X, is $E[X] = np$. Have we been

doing a Binomial experiment with Exercise 2b?

The BINOMDIST() function

BINOMDIST() computes the *probabilty* of observing k successes, in an experiment with n trials. The Inputs/Arguments are:

- `number_s` : integer
  - The number of *successes* you wish to observe.
  - Must be between 0 and the **total** number of trials.
- `trials` : integer
  - The total number of *trials*, n, that are in your experiment.
  - Must be greater than or equal to number of *successes*, k.
- `probability_s` : numeric
  - Probabiltiy of a success. Given, or empirically estimated.
  - Must be between 0 and 1.
- `cumulative` : `TRUE` or `FALSE`
  - if `FALSE` then computes the *exact* probability.
  - if `TRUE` then computes the *cumulative* probability.

Cumulative Distribution

- The Cumulative distribution is an *accumulation* of all the probablities up to and including the probability last observed.
- Show yourself how the Cumulative Distribution feature works on BINOMDIST by testing how the probabilities change when using `TRUE` in BINOMDIST versus using `FALSE`.

Exercise 3: Smokers

A local health department counsels patients coming to a clinic on cigarette smoking only if they are smokers. History has shown that about 27% (i.e., 0.27) of patients are smokes when they first come to the clinic. Assume that the clinic will see 15 patients today.

1. Graph both Binomial probability distribution and the cumulative distribution.
2. Compute the following probabilities:
   - *Exactly* 10 people are smokers.
   - 10 people or more are smokers.
   - 5 or fewer people are smokers.
   - between 7 and 10 people (inclusive) are smokers.
3. What is the *Expected* number of smokers that come into the clinic?

Exercise 3: Solution

1. Create a table that contains headers for
   - The number of smokers out of 15 patients,
   - The Binomial Probability,
   - The Cumulative Probability.
2. In Column A list the number of smokers in each case [0,15].
3. In Column B compute the Binomial Probability for each case.
   - Example: In Cell B2 write `=BINOMDIST(A2, 15, 0.27, 0)`
4. In Column C copmute the *Cumulative* Binomial Probability.
   - Example: In Cell C2 write `=BINOMDIST(A2, 15, 0.27, 1)`
5. Select Cells A1:D17 and Insert an X,Y Scatter Graph, or a Column Chart

Exercise 3: Solution (continued)

1. Compute the Probabilities by adding all the cases that satisfy the conditions in the question
   - $p(10) =$ `=B12`
   - $p(X \geq 10) =$ `=SUM(B12:B17)`
   - $p(X \leq 5) =$ `=C7`
   - $p(7 \leq X \leq 10) =$ `=SUM(B9:B12)`

	A	B	C	D	E	F	G	H
1	Smokers	Probability	Cumulative		Ranges	Numbers	itemized	Probabilities
2	0	0.00890929	0.00890929		Exactly 10 people	p(10)	p(10)	0.00128176
3	1	0.04942823	0.05833752		10 or more	p(X>=10)	p(10) + p(11) + p(12) + p(13) + p(14) + p(15)	0.00152621
4	2	0.12797172	0.18630924		5 or fewer	p(X<=5)	p(5) + p(4) + p(3) + p(2) + p(1) + p(0)	0.80418096
5	3	0.20510536	0.3914146		between 7 and 10	p(7<= X<=10)	p(7) + p(8) + p(9) + p(10)	0.08142026
6	4	0.22758266	0.61899726					
7	5	0.1851837	0.80418096					
8	6	0.11415433	0.91833529					
9	7	0.05428474	0.97262003					
10	8	0.02007792	0.99269795					
11	9	0.00577584	0.99847379					
12	10	0.00128176	0.99975555					
13	11	0.00021549	0.99997104					
14	12	2.6567E-05	0.99999761					
15	13	2.2676E-06	0.99999988					
16	14	1.1981E-07	1					
17	15	2.9543E-09	1					

The Poisson Distribution

A Poisson model follows the same idea as the Binomial Model but instead of counting k-successes in a set of n-discrete trials, it counts the number of *successes/arrivals* in a *span of time/space* in the form of a rate, successes/arrivals per unit of time/space, λ.
Suppose we are rolling a D20 die 100 times in period t seconds long. The probability of getting a Crit 20, is 1/20. for 100 rolls we Expect the rate to observe 5 Crit 20's. In a period of t seconds we expect to see 5 crits.
To approximate a Binomial Model using a Poisson Model, we can say that:

The Poisson Distribution: Continued

- In a Binomial Model we would say $p = 0.05$, $n = 100$, and $E[X] = np$.
- In a Poisson model we would say $\lambda = 5/t$ crits per second,
  - for a period length t, $E[X] = \lambda t = 5$ crits every t seconds.
- for both, we are interested in computing the probability of finding k or x successes/arrivals.

The POISSON() Function

POISSON computes the probability of x successes for a given *mean* rate of success, λ. (i.e., What is the probability two trains arrive in 10 minutes if their rate is typically 3 trains an hour) The arguments/inputs are

- `x` : integer
  - The number of arrivals/successes to be observed
  - Must be greater than or equal to 0.
- `mean` : numeric
  - The given rate of arrivals/successes.
  - Must be greater than or equal to 0.
- `cumulative` : `TRUE` or `FALSE`
  - if `FALSE` then computes the *exact* probability.
  - if `TRUE` then computes the *cumulative* probability.

Example: Counting Train Arrivals

Suppose I am watching trains at the 238$^{\text{th}}$ St station. The schedule says the train arrives every 8 minutes, for a *mean* rate of $\lambda = 0.125$ trains per minute.Use POISSON() to compute the probabilities for the number of trains, *x*, that can arrive in the next minute.

No. of Trains, *x*	Computation	Result
0	`=POISSON(0, 0.125, FALSE)`	0.8825
1	`=POISSON(1, 0.125, FALSE)`	0.1103
≥ 1	`=1-POISSON(0, 0.125, FALSE)`	0.1175

Exercise 4: Hospital Supply Quality Control

A hospital supply room manager found that on average about 2 gloves in a box are not usable.

1. Graph the probability and cumulative distributions for the Poisson distribution function with $\lambda = 2$
2. Set the probability/cumulative outputs as a number with 4 decimal places.
3. Experiment to see how many scores are needed to capture the sample space for the Poisson distribution function with $\lambda = 2$.
4. Compute the following probabilities:
    - Only 1 glove will be unusable.
    - Fewer than 5 gloves will be unusable.
    - At least 2 gloves will be unusable.

Exercise 4: Solution

1. In cells A1:C1, write the column headers ‘No. of Defects’, ‘Probability’, ‘Cumulative’.
2. Fill in integers from 0 to 10 in Cells A2:A12.
3. In cell B2, compute the probability for the given number of defects in A2.
    - Cell B2 should read: `=POISSON(A2, 2, FALSE)`
4. Use Autofill to compute the probabilities for each No. of defective gloves.
5. In Cell C2, compute the *cumulative* probability for number of defects from 0 defects up *through* the number of defects given in A2.
    - Cell C2 should read: `=POISSON(A2, 2, TRUE)`
6. Use Autofill to compute the successive cumulative probabilities in cells C3:C12.

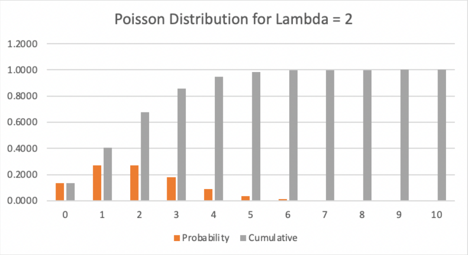
Exercise 4: Solution (continued)

1. Create a column chart of the probabilities and cumulative probabilities. Notes:
   - There should only be two series', Probability and Cumulative.
   - The 'Horizontal (Category) axis label' should be a selection of the number of defects, in this case A2:A12
2. The Probabilities are computed in the table below:

Case	Interpretation	Computation	Result
Only 1 defect	$P(X = 1)$	`=POISSON(1, 2, FALSE)` or See Cell B3	0.2707
Fewer than 5 defects	$P(X < 5)$ or $P(X \leq 4)$	`=POISSON(4, 2, TRUE)` or See Cell C6	0.9473
At least 2 defects	$P(X \geq 2)$ or $1 - P(X \leq 1)$	`=1-POISSON(1, 2, TRUE)` or =1-C3	0.5940

Exercise 4: Solution (continued)

	A	B	C	D	E	F	G	H
1	No. of Defects	Probability	Cumulative		Case/Defects	Interpretation	Itemized	Number
2	0	0.1353	0.1353		Only 1	P(X=1)	p(1)	0.2707
3	1	0.2707	0.4060		Fewer than 5	P(X<5)	P(0) + P(1) + P(2) + P(3) + P(4)	0.9473
4	2	0.2707	0.6767		At Least 2	P(X>=2)	P(2) + P(2) + P(3) + ...	0.5940
5	3	0.1804	0.8571					
6	4	0.0902	0.9473					
7	5	0.0361	0.9834					
8	6	0.0120	0.9955					
9	7	0.0034	0.9989					
10	8	0.0009	0.9998					
11	9	0.0002	1.0000					
12	10	0.0000	1.0000					



Further Reading

The topics covered in the lecture can be found in *Compter Applications for Life Sciences* p. 63 - 74.