COMP 1433: Introduction to Data Analytics &
COMP 1003: Statistical Tools and Applications

Tutorial 2
Probability Basics

Yu Erxin
Department of Computing
The Hong Kong Polytechnic University

Jan, 2023

Roadmap

- Sample Space
- Conditional Probability and Multiplication Rule
- Bayes Theorem
- Discrete Random Variables

Sample Space

Consider two datasets below:

Dataset1:

1. @AlyssaNoelleD I knw lol imma make a picture for you
2. @alex_gibson lacks features, but three words... fast, fast, fast !
3. @catchthesunx it should I reckon ! I wanna goooo ! LET'S GO CRAZY

Dataset2:

1. A historical epic with the courage of its convictions about both scope and detail .
2. A fun family movie that's suitable for all ages -- a movie that will make you laugh , cry and realize , 'it's never too late to believe in your dreams .
3. Renner carries much of the film with a creepy and dead-on performance .

Questions

1.You are required to calculate the word frequency w.r.t. each dataset. What are the sample spaces? Give some examples of word probabilities.
2.Can you guess the source(s) of these two datasets? And why?
3.What else can you find in these datasets?

Sample Space

How to calculate word frequency?

Example: “I have told you I am the best.”

Steps 1. total word count = 9

2. word frequency:

‘I’:2	‘have’:1	‘told’: 1	‘you’: 1
‘am’: 1	‘the’: 1	‘best’: 1	‘.’: 1

3. word probability:

‘I’: 2/9	‘have’:1/9	‘told’: 1/9	‘you’: 1/9
‘am’: 1/9	‘the’: 1/9	‘best’: 1/9	‘.’: 1/9

Question:

1. You are required to calculate the word frequency w.r.t. each dataset. What are the sample spaces? Give some examples of word probabilities.
2. Can you guess the source(s) of these two datasets? And why?
3. What else can you find in these datasets?

Sample Space

Question:

1. You are required to calculate the word frequency w.r.t. each dataset. What are the sample spaces? Give some examples of word probabilities.

   The sample space of an experiment is all the possible outcomes for that experiment.

   The vocabulary of each dataset.

   Dataset1 has 37 words. P (X = ‘fast’) = 3/37, P(X = ‘@’) = 3/37

2. Can you guess the source(s) of these two datasets? And why?

   1st dataset is extracted from Twitter/Facebook or other social media website.

   2nd dataset is extracted from IMDB or other movie review website.

   Some evidence: probability of ‘@’, misspelling, gramma mistakes, etc.

3. What else can you find in these datasets?

   The writing style of 1st dataset is less formal than 2nd dataset.

   More grammatical incompleteness. More slang. Higher probability of ‘@’, etc.

Conditional probability and the multiplication rule

Question:

Two cards are dealt from a deck of 52 cards. What is the probability that they are both Aces? And how about the cases of three and four cards which are all Aces?

Conditional probability and the multiplication rule

Question:

Two cards are dealt from a deck of 52 cards. What is the probability that they are both Aces? And how about the cases of three and four cards which are all Aces?

Solution:

Let's use conditional probability. Let A1 be the event "first card an Ace" and A2 be the event "second card an Ace". P(A2|A1) is the probability of a second Ace. Given that the first card has been drawn and was an Ace, there are 51 cards left, 3 of which are Aces, so P(A2|A1) = 3/51.

Then we use the Multiplication Rule. the probability of occurrence of both events A1 and A2 is equal to the product of the probability of A1 occurring and the conditional probability that event A2 occurring given that event A1 occurs.

$$\begin{aligned} \mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_1) \times \mathrm{P}(A_2|A_1) \\ &= \frac{4}{52} \times \frac{3}{51} \\ &= \frac{1}{221}. \end{aligned}$$

Bayes Theorem

Question:

A veterinary offers you a free test for a very rare (affecting only one in ten thousand), but hideous disease for your lovely pet. The test they offer is very reliable. If your pet has the disease it has a 98% chance of giving a positive result, and if not, it has only a 1% chance of giving a positive result. You decide to take the test, and find that your pet test positive — what is the probability that your pet really have the disease?

Bayes Theorem

Question:

A veterinary offers you a free test for a very rare (affecting only one in ten thousand), but hideous disease for your lovely pet. The test they offer is very reliable. If your pet has the disease it has a 98% chance of giving a positive result, and if not, it has only a 1% chance of giving a positive result. You decide to take the test, and find that your pet test positive — what is the probability that your pet really have the disease?

Bayes Theorem

Let P be the event “test positive” and D be the event “your pet has the disease”. Then

If your pet has the disease it has a 98% chance of giving a positive result $\mathrm{P}(P|D) = 0.98$

if not, it has only a 1% chance of giving a positive result $\mathrm{P}(P|D^c) = 0.01$

We want to know $P(D|P)$, so we use Bayes’ Theorem:

$$\begin{aligned}\mathrm{P}(D|P) &= \frac{\mathrm{P}(P|D)\ \mathrm{P}(D)}{\mathrm{P}(P)} \\ &= \frac{\mathrm{P}(P|D)\ \mathrm{P}(D)}{\mathrm{P}(P|D)\ \mathrm{P}(D) + \mathrm{P}(P|D^c)\ \mathrm{P}(D^c)} \quad \text{(using the theorem of total probability)} \\ &= \frac{0.98\,\mathrm{P}(D)}{0.98\,\mathrm{P}(D) + 0.01(1 - \mathrm{P}(D))}.\end{aligned}$$

$P(D) = 0.0001$. (affecting only one in ten thousand)

$$\mathrm{P}(D|P) = \frac{0.98 \times 0.0001}{0.98 \times 0.0001 + 0.01 \times 0.9999} \simeq 0.01.$$

Bayes Theorem

Tips:

1. Be cautious about modelling the correct conditional probability from question.
2. Note the crucial differences between P(P|D) and P(D|P).
3. Note the crucial differences between P(D) and P(D|P). In this question, the probability of your pet having the disease has increased from 0.0001 to 0.01. But still isn't that much to get worried about!
4. P(D) and P(D|P) are often called prior probability and posterior probability.

Discrete Random Variables

Roll one dice and call the random number which is uppermost *Y*. Roll two dice and call their sum *Z*.

Question:

1.What are the sample spaces of Y and Z? Are the outcomes of Y equally likely? How about Z?

2.What is the Probability mass function of Z? (Probability mass function is the probability of a discrete random variable on each specific value)

3. You are rolling two dice with your friend and the larger sum of dice shall win. What is the chance you beat him/her when he/she rolls (4 and 5)?

Discrete Random Variables

Roll one die and call the random number which is uppermost *Y*. Roll two dice and call their sum *Z*.

Solution:

1. What are the sample spaces of Y and Z? Are the outcomes of Y equally likely? How about Z?

$$S_Y = \{1, 2, 3, 4, 5, 6\}$$

$$S_Z = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$ But never 1

Discrete Random Variables

Roll one die and call the random number which is uppermost *Y*. Roll two dice and call their sum *Z*.

Solution:

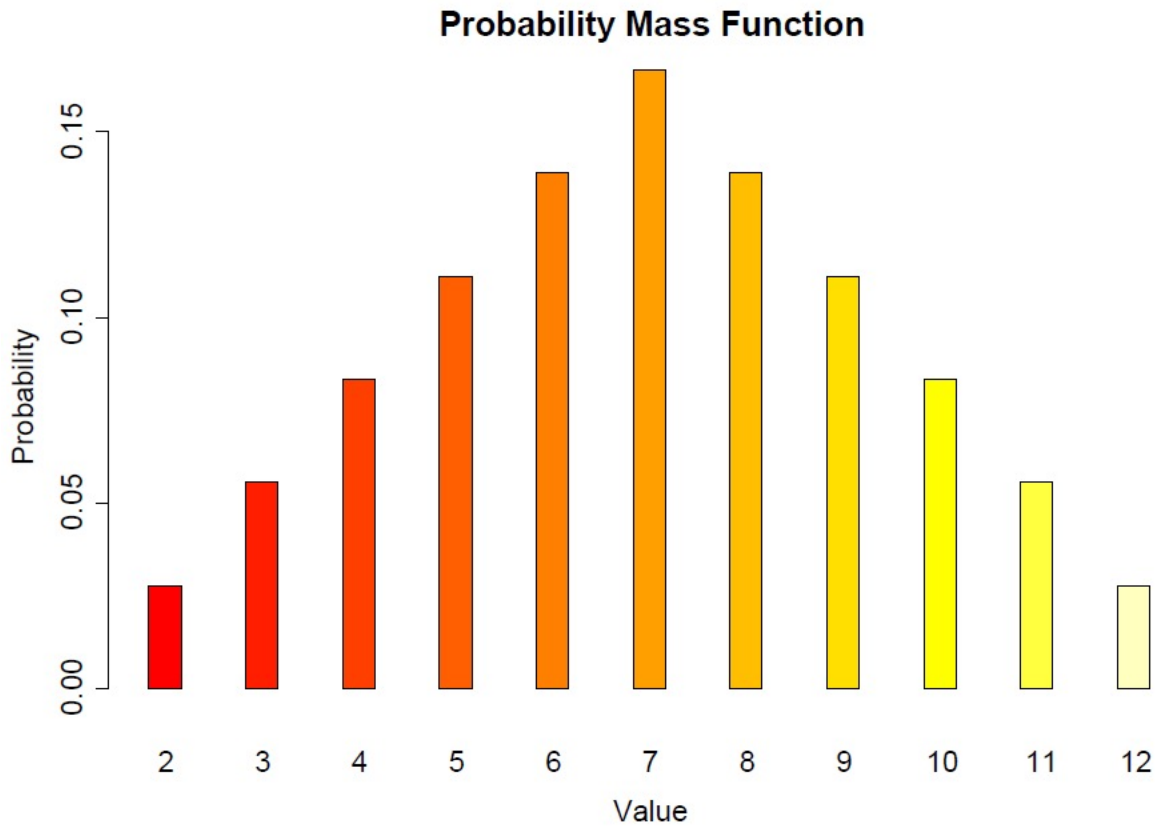
2. What is the PMF of Z?

Probability mass function can be tabulated as

x	2	3	4	5	6	7	8	9	10	11	12
$\mathrm{P}(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Discrete Random Variables

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



Discrete Random Variables

Roll one dice and call the random number which is uppermost *Y*. Roll two dice and call their sum *Z*.

Question:
3. You are rolling two dice with your friend and the larger sum of dice shall win. What is the chance you beat him/her when he/she rolls (4 and 5)?

Solution:
A. $P(X>4+5) = P(X=10) + P(X=11) + P(X=12) = 3/36 + 2/36 + 1/36 = 1/6$
B. $P(X>4+5)=1–P(X<=4+5)=1-F_X(9)=1–30/36=1/6$