7. Fundamentals of Probability

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

Summary:

1. Control structures
   - Conditionals: `if`, `if-else`, `ifelse`
   - Iterators: `for`, `while`, `repeat`
2. Functions

Maintain a glossary of functions used.

Congratulations!

This completes the first part of the course

An Introduction to R programming for Data science

Next we will look at

Introductory Probability and Statistics in R.

Some calculations for probability will be done by hand(using paper and pencil)

Next we will see...

- Probability
  - Simulation approach to probability
  - Basic Probability Definitions
  - Probability Properties and rules: complement, multiplication, Addition,
  - Independent events, Conditional Probability
  - Mutually exclusive events
  - Practice is key.

Simulation

- Uncertainty
- Chances

Basic Definitions and Examples

Experiment, sample space, event

- An **experiment** is any activity for which the outcome is uncertain.
- A **random experiment** is one in which we know all the **possible outcomes** in advance but we do not know which outcome will occur when the experiment unfolds.
- The set of all these possible outcomes is called the **sample space.**
- An **event** is a subset of the sample space.

Example

When tossing a coin, there are two possible outcomes, “head” or “tail,” and we choose to be interested in getting a head.

Experiment: Toss a coin

Sample space: $S = \{H, T\}$

Event: $E = \{H\}$

Example

We roll a die and look for an even number.

Example

We roll a die and look for an even number.

Experiment: Roll a die

Sample space: $S = \{ 1, 2, 3, 4, 5, 6\}$ (the six faces)

Event: $E = \{2, 4, 6\}$

Some other events:

- $A = \{4\}$ (rolling a 4)
- $B = \{1, 3, 5\}$ (rolling an odd number)

Example

We can roll a six sided die twice and look for a double.

Example

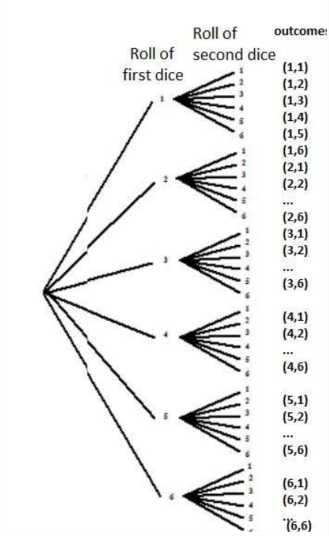
We can roll a six sided die twice and look for a double.

Experiment: Roll two dice

Sample space: S = $\{(1, 1), (1, 2), ... , (1, 6),$
$(2, 1), . . . ,. . ., . .(2,6),$
$(6,1). . . ,. . . ,. . . .. (6, 6)\}$ (36 outcomes)

Event: E = $\{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

Sample space for roll of two dice: Tree diagram



Classical Approach to Probability

If the number of outcomes is finite and they are **all equally likely to occur**, then

$$\mathbb{P}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of possible outcomes}}$$

Example

- **Experiment:** Roll a fair six sided die once.
- **Sample space:** $\{1, 2, 3, 4, 5, 6\}$
- Find the probability of scoring a 4.
  - Let Event A = scoring a 4 =$\{4\}$
  - $\mathbb{P}(A) = 1/6$.

Example

What's the probability of getting a double when rolling two dice?

- **Experiment:** Roll two dice
- **Sample space:** S $= \{(1, 1), (1, 2), \ldots , (1, 6),$
  $(2, 1), \ldots ,\ldots ., \ldots (2,6),$
  $(6,1)\ldots,\ldots,\ldots.. (6, 6)\}$ (36 outcomes)
- **Event:** E $= \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$
- Find the probability of getting a double.
  - Let Event D = getting a double
  - $\mathbb{P}(D) = 6/36 = 1/6.$

Relative frequency approach to probability

1. repeat the relevant experiment over and over again, say n times
2. count how many times, say e times, the event E occurs in these n reps,
3. The "relative frequency" of the event E is the proportion e/n
4. The *probability* of E is this proportion e/n when we take the number of reps , $n \to \infty$.

We can simulate the relative frequency approach in `R` using the `sample` function

Certain & Impossible Events and probabilities

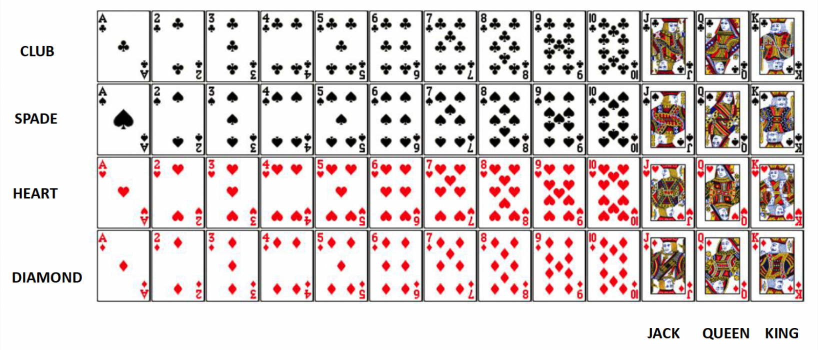
The probability of an event that is **CERTAIN** to occur is 1.

The probability of an **IMPOSSIBLE** event is 0.

- **Experiment:** Roll a fair six sided die once.
- **Sample space:** $\{1, 2, 3, 4, 5, 6\}$
- Let Event A = scoring a 1, 2, 3, 4, 5, OR 6 $= \{1, 2, 3, 4, 5, 6\}$
  - $\mathbb{P}(A) = 6/6 = 1$.
- Let Event B be the event of 'rolling a 7' when a fair six sided die is rolled; then $P(B) = 0/6 = 0$.

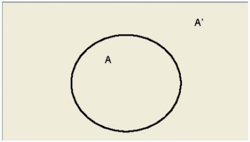
The probability of any event A, $P(A)$, **is a number between 0 and 1**; $0 \leq P(A) \leq 1$

Standard Deck of Cards



Complement of an event

The **complement** of event A, denoted by (A^c or by $\bar{A}$ or A'), is the set of outcomes in the sample space S, that are not included in the outcomes of event A.



Example When drawing a card from a deck, if A consists of an ace, then A' consists of all those cards that are not aces.

Probability of NOT getting an ace is
$P(A') = 1 - P(A) = 1 - 4/52 = 48/52$

Independent Events

DEFINITION: Two events are *independent* if the occurrence of one of the events does not affect the chance(probability) that the other event occurs.

Two events are dependent if the chance of the second event changes depending on whether or not the first event happened.

Example Drawing Cards from a Deck with and without replacement.

- with replacement : independent
- without replacement: dependent.

Are these independent or dependent

- Rolling a fair die twice
- Guessing on each answer for a 10-question True/False test
- Flipping a coin 4 times
- Drawing students to take a survey randomly (without replacement)

Multi-event probabilities: Multiplication Rule

The multiplication rule is used to calculate the probability of two things happening at the same time.

Multiplication Rule (Independent Events) If we have two *independent events* then the multiplication rule is:

$$P(\text{A and B}) = P(A) \times P(B)$$

Example: Rolling Two Dice

The probability of rolling two dice and getting the first to be "1" and the second to be "2"

$$P(\text{1 and then a 2}) = P(1) * P(2) = 1/6 * 1/6 = 1/36$$

Multiplication Rule (Dependent Events)

For dependent events, the multiplication rule is

$P(\text{A and B})) = P(A)P(B|A)$, where $P(B|A)$ is the probability of event B given that event A happened.

Example: Drawing Two Cards without Replacement

The probability of drawing 2 cards without replacement from a deck and getting a heart and then a spade can be found using the Multiplication Rule.

$P(\text{heart and then spade})) = P(\text{heart})P(\text{spade}|\text{heart}) = 13/52 * 13/52$

Example:

- Draw 2 cards with replacement. What's the chance of getting two red cards?

- Draw 2 cards without replacement. What's the chance of getting two red cards?

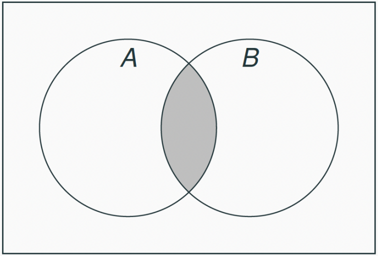
- Draw 2 cards without replacement. What's the chance of getting the first red and second not red? Becaues of dependence,

Example:

- Draw 2 cards with replacement. What's the chance of getting two red cards? Because of independence, $P(\text{red and then red})) = P(\text{red})P(\text{red}) = 26/52 * 26/52 = (26/52)^2 = (1/2)^2 = 1/4 = 0.25$
- Draw 2 cards without replacement. What's the chance of getting two red cards? Becaues of dependence $P(\text{red and then red})) = P(\text{red})P(\text{red given first was a red}) = 26/52 * 25/51 = 0.245$
- Draw 2 cards without replacement. What's the chance of getting the first red and second not red? Becaues of dependence, $P(\text{red and then NOT red})) = P(\text{red})P(\text{NOT red given first was a red}) = 26/52 * 26/51 = 0.255$

Intersection of events, AND

In general, if A and B are events in the sample space, then the intersection $A \cap B$ denotes the outcomes in "A and B".



Draw one card from a standard deck of 52 cards,

- A = Spade = { AS, 2S, ... KS },
- B = Ace = {AS, AC, AD, AH }
- $A \cap B$ = Spade and Ace = {AS }
- $P(A \cap B) = P(\text{A and B}) = 1/52$

May seem complicated but just need to remember the formula and not list outcomes:

- $P(A \cap B) = P(\text{Spade})P(\text{Ace given Spade}) = 13/52 * 1/13 = 1/52$
- $P(A \cap B) = P(\text{Ace})P(\text{Space given Ace}) = 4/52 * 1/4 = 1/52$

Multi-event probabilities: Addition Rule

The Addition Rule is used to calculate the probability that either (or both) of 2 events will happen

$P(\text{A or B})) = P(A) + P(B) - P(\text{A and B})$, where $P(\text{A and B})$ is subtracted to avoid double counting

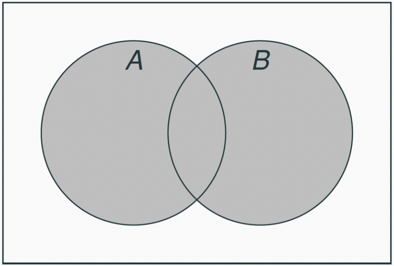
Example:

What's the probability of drawing either a queen or a heart from a standard deck of cards?

$P(\text{queen OR heart})) = P(\text{queen}) + P(\text{heart}) - P(\text{queen AND heart}) = 4/52 + 13/52 - 1/52 = 16/52 = 0.3077$

Union, OR

If A and B are events in the sample space, then the union $A \cup B$ denotes the outcomes in "A or B".



Draw one card from a standard deck of 52 cards,

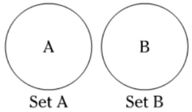
- A = Spade = { AS, 2S, ... KS },
- B = Ace = {AS, AC, AD, AH }
- $A \cup B$ = Spade or Ace = {AS, 2S, ... KS, AC, AD, AH }
- $P(A \cup B) = 16/52$

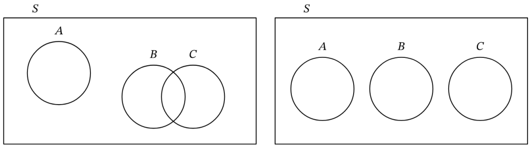
May seem complicated but just need to remember the formula and not list outcomes:

- $P(A \cup B) = P(\text{Spade OR Ace}) = P(\text{Spade}) + P(\text{Ace}) - P(\text{Spade AND Ace})) = 13/52 + 4/52 - 1/52 = 16/52$

MUTUALLY EXCLUSIVE (DISJOINT) EVENTS

Two events A and B are said to be mutually exclusive if they cannot occur together.





A is mutually exclusive to *B* and *C*, but *B* and *C* are not mutually exclusive.

A, *B* and *C* are pairwise mutually exclusive.

A and B are mutually exclusive then A and B do not share any outcomes, they are non-overlapping.

mutually exclusive or disjoint

For example, if we pull a card from a deck and consider A = Spade, B = Heart

With one card selected, it is impossible to get both a heart and a spade; we may get one or the other but not both.

$A \cap B = \emptyset$(the empty set) and the joint probability is zero: $P(A \cap B) = 0$

A and B are mutually exclusive or disjoint.

non mutually exclusive

A = Spade, B = Ace then $A \cap B = \{\text{AS}\}$

$P(A \cap B) = 1/52$

So, $P(A \cap B) \neq 0$ and A and B are not mutually exclusiv.

A and B are not mutually exclusive means they share some outcomes. ie they are overlapping.

Special case: Multi-event probabilities: Addition Rule

Recall, The Addition Rule is used to calculate the probability that either (or both) of 2 events will happen

$P(\text{A or B})) = P(A) + P(B) - P(\text{A and B})$, where $P(\text{A and B})$ is subtracted to avoid double counting

Special case: When A and B are mutually exclusive, the addition rule simplifies to

$P(\text{A or B}) = P(A) + P(B)$

Summary: Probability Properites

1. The probability of an event A, denoted by $P(A)$, is a number between 0 and 1. $0 \leq P(A) \leq 1$
2. For the sample space S, P(S) = 1
3. $P(\emptyset) = 0$
4. Complement: $P(A^c) = 1 - P(A)$
5. addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   - **Special case:** If A and B are mutually exclusive events, that is, $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$
6. multiplication rule: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$, where $P(B|A)$ is the probability of event B given that event A happened.
   - **Special case:**If A and B are independent, ie $P(B|A) = P(B)$, then $P(A \cap B) = P(A) \times P(B)$

Example: Conditional Probability

The table shows the results of a study in which researchers examined a child's IQ and the presence of a specific gene in the child.

	Gene Present	Gene Absent	Total
High IQ	33	19	52
Normal IQ	39	11	50
Total	72	30	102

Find the probability that…

(i) … a child has a high IQ, given the child has the gene.
(ii) … a child has the gene.
(iii) … a child has a high IQ, and the child has the gene.

Let Event A = Gene Present

Let Event B = High IQ

(i) P(a child has high IQ, **given** that the child has the gene) =

$P(B \mid A) = \boxed{\dfrac{33}{72}}$,

by the Classical Definition.

(ii) P(a child has the gene) $= P(A) = \boxed{\dfrac{72}{102}}$

(iii) P(a child has high IQ, **and** the child has the gene) =

$P(B \text{ and } A) = \boxed{\dfrac{33}{102}}$

So:

$$\frac{P(B \text{ and } A)}{P(A)} = \frac{33}{102} \Big/ \frac{72}{102} = \frac{33}{72} = P(B \mid A)$$

Definition of Conditional Probability:

Given A has already occured, the probability of B given A is

$$P(B \mid A) = \frac{P(B \text{ and } A)}{P(A)} = \frac{P(B \cap A)}{P(A)}, \quad \text{when } P(A) > 0$$

Multiplication Law of Probability:

$$P(B \cap A) = P(A)P(B|A)$$

Your Turn

CD's in a music shop are classified as: classical, pop, rock, folk and jazz. The probability that a customer buying one CD will choose classical is 0.3, pop 0.4, rock 0.2, folk 0.05 and jazz 0.05

a) Find the probability that a customer will choose a classical, folk or jazz CD.
b) Find the probability that a customer will NOT choose a classical, folk or rock CD.

Assume each CD can only be classified in one section

Outline Solutions

a) $P(\text{classical OR folk OR jazz}) = 0.3 + 0.05 + 0.05 = \boxed{0.4}$

b) $P(\text{classical OR folk OR } \textbf{rock}) = 0.55$; Therefore the probability that a customer will NOT choose a classical, folk or rock CD is $1 - 0.55 = \boxed{0.45}$.

Your Turn

A fair six sided die is thrown 4 times. Find the probability that a 5 is obtained each time.

STRATEGY:

- Are these events independent? multiply probabilities or use multiplication rule
- Are these events mutually exclusive? add probabilities or use addition rule
- If the number of ways your event can unfold is long and complicated, look at the complemet.

Once you have worked this out, apply the correct formula.

Outline Solutions

Let A_i denote the event 'the i^{Th} roll resulted in the number 5', for $i = 1, 2, 3, 4$. These events are **NOT mutually exclusive** (it *is* possible to obtain a 5 on the, say, first and third rolls);

the events are **independent**. Therefore,

$$\begin{aligned} &P(A_1 \text{ AND } A_2 \text{ AND } A_3 \text{ AND } A_4) \\ &= P(A_1) \times P(A_2) \times P(A_3) \times P(A_4) \\ &= \boxed{\left(\frac{1}{6}\right)^4} \end{aligned}$$

Probability Problem

A PSTAT midterm exam consists of multiple choice questions. Each question has 4 possible answers. Based on your performance in the class, you decide that your probability ***knowing*** the correct answer to any question is 0.75. If you do ***not know*** the correct answer, you intend to guess.

What is the probability you will choose the correct answer to a question?

- Let A be the event that you give the correct answer.
- Let B be the event that you knew the correct answer. (0.75)
- We want to find $P(A)$.
- $P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$

- $P(A \text{ and } B) = P(A \mid B) \cdot P(B) = 1 \times 0.75 = 0.75$
- $P(A \text{ and } B^c) = P(A \mid B^c) \cdot P(B^c) = 0.25 \times 0.25 = 0.0625$. Note: you have a 1 in 4 chance of choosing the correct answer randomly
- So, $P(A) = 0.75 + 0.0625 = \boxed{0.8}$

Law of total probability $P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$

–>

We did:

- Probability
  - Definitions
  - Rules of Probability: Addition, complement, multiplication
  - Conditional Probability
  - Mutually exclusive events
  - Independent events