

7. Fundamentals of Probability

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

->

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

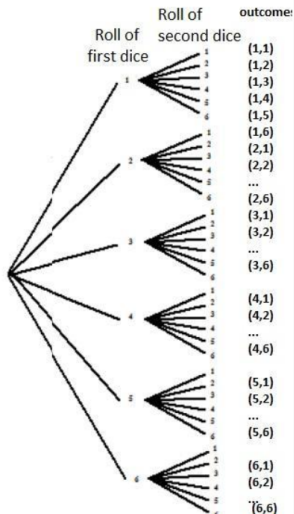
We can roll a six sided die twice and look for a double.

Experiment: Roll two dice

Sample space: $S = \{(1, 1), (1, 2), \dots, (1, 6),$
 $(2, 1), \dots, \dots, \dots (2, 6),$
 $(6, 1), \dots, \dots, \dots (6, 6)\}$ (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

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

- **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** $A = \text{scoring a } 4 = \{4\}$
 - $\mathbb{P}(A) = 1/6$.

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 , $n \rightarrow \infty$.

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

Simulating probabilities in R

Example: Toss a coin

```
(x<-sample( c("H", "T"), 10, replace = TRUE)) #10 coin tosses
```

```
## [1] "H" "T" "H" "T" "T" "T" "H" "H" "H" "H"
```

```
table(x) # How many H vs T
```

```
## x
```

```
## H T
```

```
## 5 5
```

```
table(x)/10 #Estimated probabilities
```

```
## x
```

```
## H T
```

```
## 0.2 0.8
```

What happens with 100 coin tosses, 1000 coin tosses?

```
x<-sample( c("H", "T"), 100, replace = TRUE)
table(x)/100
```

```
## x
##   H   T
## 0.47 0.53
```

```
x<-sample( c("H", "T"), 1000, replace = TRUE)
table(x)/1000
```

```
## x
##   H   T
## 0.51 0.49
```

Example 2

A manufacturing process produces 20% defective items. Simulate this process.

First Decide what's the experiment, and what are the possible outcomes (sample space) of this experiment?

Then repeat the experiment many times by appropriate use of sample

Example 2

A manufacturing process produces 20% defective items. Simulate this process.

What's the experiment, and what are the possible outcomes (sample space) of this experiment? **Experiment** Producing an item

Sample space $S = \{ d, g \}$

Simulate production of 10 items from this process.

```
p <- sample( c("d","g"), 10 , replace = TRUE,  
            prob = c(0.2, 0.8))
```

```
table(p)/10
```

```
## p
```

```
##   d   g
```

```
## 0.1 0.9
```

Example

Ethan and Adam are gambling against each other. A fair coin is tossed repeatedly. Each time a head comes up, Ethan wins a \$ from Amy, and each time a tail comes up, Ethan loses a \$ to Amy. Carry out this experiment 50 times, and estimate the number of times that Ethan is ahead in these 50 tosses. How much has Ethan won or lost?

1. generate 50 tosses of a fair coin

```
x <- sample(c("H", "T") , 50, replace = TRUE )
```

equivalently, from Ethan's perspective

```
x <- sample(c(1, -1) , 50, replace = TRUE )
```

```
x  
  
## [1] -1 -1 -1 1 -1 1 -1 -1 -1 -1 -1 -1 1 1 1 1 1 -1  
## [26] 1 -1 1 1 -1 -1 1 -1 1 -1 -1 -1 1 1 -1 1 -1
```

Ethan's winnings

```
x[1] # winning after 1st toss
```

```
## [1] -1
```

```
x[1]+x[2]
```

```
## [1] -2
```

```
x[1]+x[2]+x[3]
```

```
## [1] -3
```

```
sum <- rep(0,50)
```

```
sum[1] <- x[1] #the winnings after the first toss
```

```
for (i in 2 :50) {
```

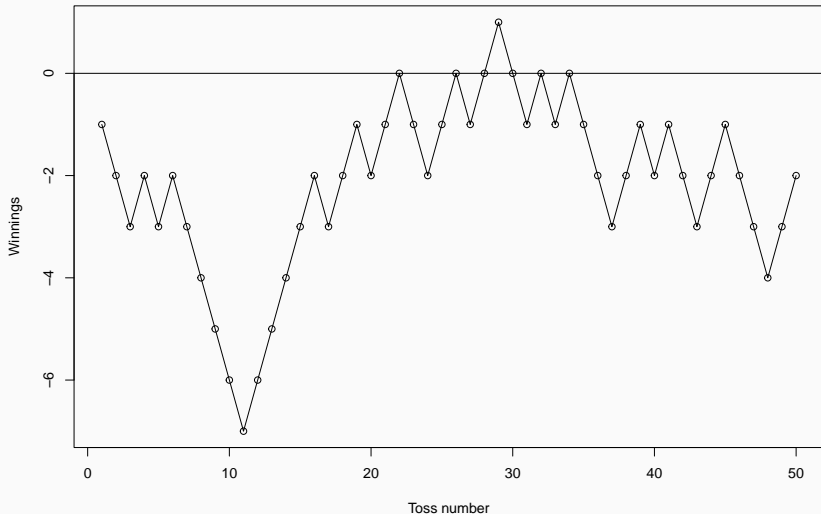
```
  sum[i] <- sum[i-1] + x[i] #the winnings at toss i
```

```
}
```

```
sum
```

```
## [1] -1 -2 -3 -2 -3 -2 -3 -4 -5 -6 -7 -6 -5 -4 -3 -2 -3  
## [26]  0 -1  0  1  0 -1  0 -1  0 -1 -2 -3 -2 -1 -2 -1 -2
```

```
plot(1:50, sum, type = "o",  
     xlab = "Toss number", ylab = "Winnings")  
abline( 0 , 0 )
```



overall winning





















































```
sum[50]
```

```
## [1] -2
```

```
sum(x)
```

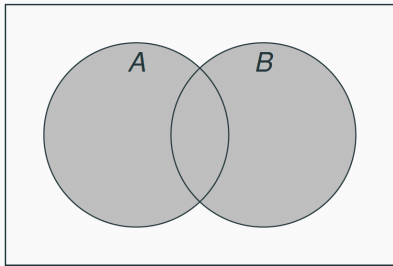
```
## [1] -2
```

PROBABILITY AND SETS

CLUB													
SPADE													
HEART													
DIAMOND													
											JACK	QUEEN	KING

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 ”.

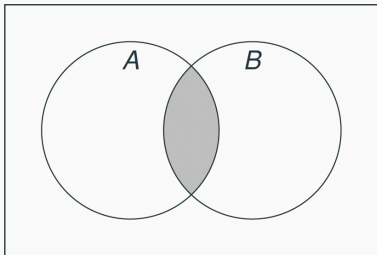


Draw one card from a standard deck of 52 cards,

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

Intersection, AND

If A and B are events in the sample space, then the intersection $A \cap B$ denotes the outcomes in “ A and B ”.

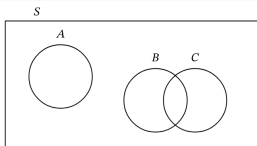
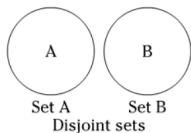


Draw one card from a standard deck of 52 cards,

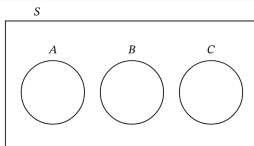
- $A = \text{Spade} = \{ AS, 2S, \dots KS \},$
- $B = \text{Ace} = \{ AS, AC, AD, AH \}$
- $A \cap B = \text{Spade and Ace} = \{ AS \}$
- $P(A \cap B) = 1/52$

MUTUALLY EXCLUSIVE 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

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

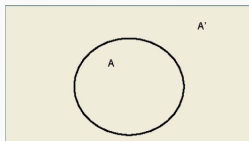
$$P(A \cap B) = 1/52$$

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.

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 S , that are not in A .



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 rules/axioms

A probability function must satisfy:

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. If A and B are mutually exclusive events, that is, $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

Probability properties

- a. complement: $P(A^c) = 1 - P(A)$
- b. $P(\emptyset) = 0$
- c. addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

A card is drawn from a well-shuffled deck. What is the probability that the card is an ace or a heart?

A = ace, H = heart

$$P(A) = 4/52, \quad P(H) = 13/52, \quad P(A \cap H) = 1/52$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/52 + 13/52 - 1/52 = 16/52$$

Certain & Impossible Events

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 $P(B) = 0/6 = 0$.

DEFINITION: A conditional probability is the probability of an event occurring **given** that another event has already occurred.

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) $P(\text{a child has high IQ, given that the child has the gene}) =$

$$P(B | A) = \boxed{\frac{33}{72}},$$

by the Classical Definition.

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

(iii) $P(\text{a child has high IQ, and the child has the gene}) =$

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

So:

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

Definition of Conditional Probability:

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

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

Multiplication Law of Probability:

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

Independent Events

DEFINITION: Two events are *independent* if the occurrence of one of the events does not affect the probability of the occurrence of the other event.

A and B are independent if

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

Example When tossing a fair coin, the probability of a head is $1/2$ in every toss; so the probability of a second head, given that the first toss is a head, remains at $1/2$.

$$P(H_2|H_1) = P(H_2) = 0.5$$

If A and B are independent then (multiplication law reduces to)

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

More than 2 Events

- If events A , B and C are **mutually exclusive** then:

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

This can be extended to any number of mutually exclusive events.

- If events A , B and C are **independent** then:

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

This can be extended to any number of independent events.

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

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

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

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

STRATEGY:

- Are these events mutually exclusive? add probabilities
- Are these events independent? multiply probabilities

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,

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

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 $P(A)$.
- $P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$

- $P(A \text{ and } B) = P(A | B) \cdot P(B) = 1 \times 0.75 = 0.75$
- $P(A \text{ and } B^c) = P(A | B^c) \cdot P(B^c) = 0.25 \times 0.25 = 0.0625$.
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 $P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$

→

Summary:

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