Week 7: Introduction to Probability

Data Analysis for Psychology in R 1

Marju Kaps

Department of Psychology
The University of Edinburgh

Brief Introduction

- Marju Kaps
- Delivering all lectures through the remainder of S1
- Contact Email: mkaps@ed.ac.uk
- Office Hours: TBD
- Office Location: 7 George Square, UG44

Learning objectives

- Understand the link between probability, models and data
- Be able to connect probability to statistical inference

Today

- Intro to Probability
- Sets & Set Notation
- Random Experiments

Part 1: Intro to Probability

Why probability?

- When conducting psychological research, we often ask a question and gather data in an attempt to identify the true answer, AKA the **ground truth**
- We want to use our data to build a model of the world
  - **Model:** a formal representation of a system
  - Put another way, a model is an idea about the way the world is
- Two types of models you could use:
  - Deterministic
  - Probabilistic/Stochastic

Why probability?

- Imagine you live exactly 1/2 mile from the building and your walking speed is 3.3 miles per hour. You want to compute how long it takes to get to class.
- You can use a deterministic model to calculate this:

- $\frac{distance}{speed} = time$
- $\frac{0.5\ miles}{3.3\ mph} = 0.15\ hours$
- $0.15 * 60 = 9\ minutes$
- Using this model, you should always be right on time as long as you leave at 8:51

But what if...

- you text while walking?
- you stop to chat with someone?
- You get stuck at an intersection for longer than normal?
- you are so excited about learning statistics that you walk especially quickly?

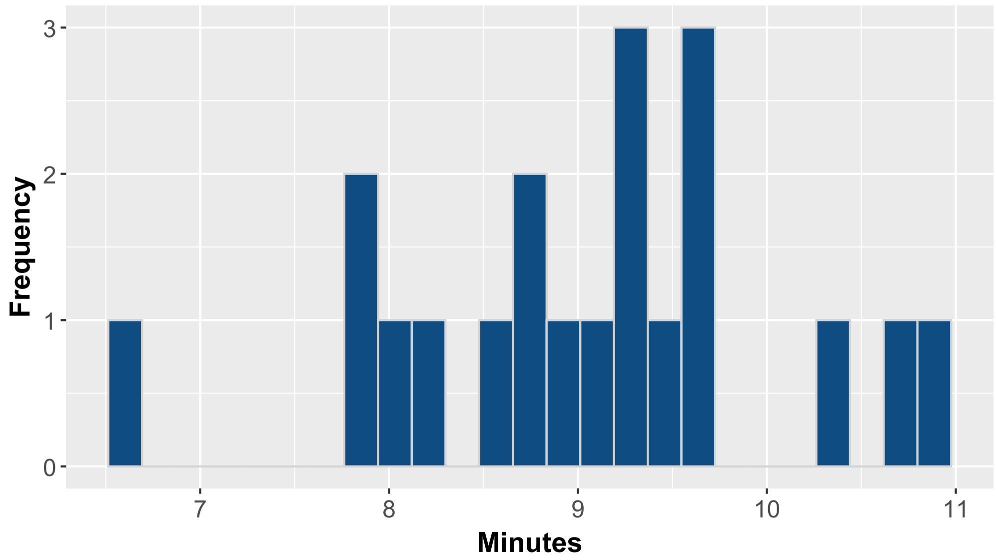
- Suddenly, your deterministic solution isn't such a great model for the world.

Why probability?

- Deterministic models imply certainty and consistency, but real world data (especially human subjects data!) are complex
  - There are many factors that we can't anticipate or account for in our studies.
- With inferential statistics, we make sense of the world using probabilistic models, which take the element of randomness into account.
- Inferential tests tell you something about the probability of your data and this helps guide your decision about the ground truth.

Why probability?

- Imagine that you timed your walk to class over the course of a month.
- These data indicate that by leaving at 8:51, the likelihood you will arrive on time is only about 45%.



Questions

What is probability?

- Likelihood of event’s occurrence
- There are two ways to conceptualise probability:
    1. Analytic Definition
    2. Relative Frequency

Probability: Analytic Definition

- The probability of an event is equal to the ratio of successful outcomes to all possible outcomes

$$P(x) = \frac{a}{a+b}$$

a = ways that event x can occur

b = ways that event x can fail to occur

x = Drawing a black card

a = # of black cards

b = # of red cards

$P(x) = \frac{26}{26+26}$

$P(x) = \frac{1}{2} = 0.50$

x = Drawing a spade

a = # of spades

b = # of diamonds + hearts + clubs

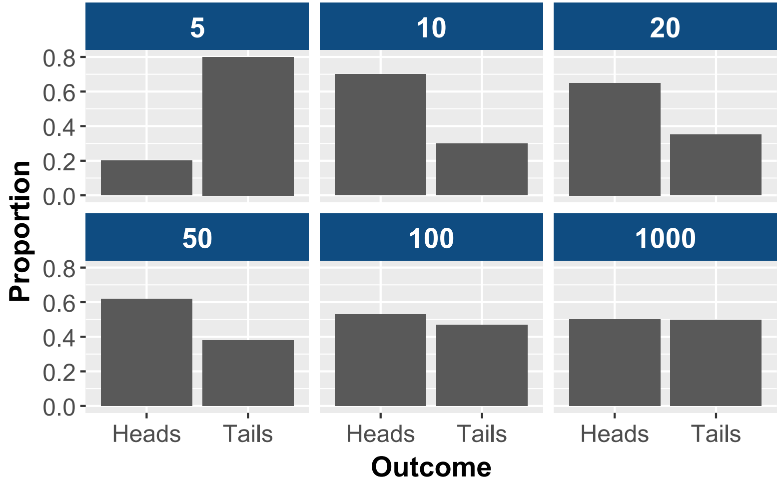
$P(x) = \frac{13}{13+39}$

$P(x) = \frac{13}{52} = 0.25$

Probability: Relative Frequency

- $P(x)$, or probability of x, is the proportion of times you would observe x if you took an infinite number of samples.
    - If I roll a die an infinite number of times, the probability I would roll a 4 would be exactly 1/6.
- **The law of large numbers**
    - Given an event x and a probability $P(x)$, over n trials, the probability that the relative frequency of x' will differ from $P(x)$ approaches 0 as n approaches infinity

x = A flipped coin landing on heads



What is probability?

- What does all that stuff get us?
- Basic idea: $P(x)$ = the number of ways x can happen divided by the number of possible outcomes (including x)
- In its most essential sense, the business of probability is figuring out the values of those two numbers
- Generally, we base these calculations on structures called *sets*...

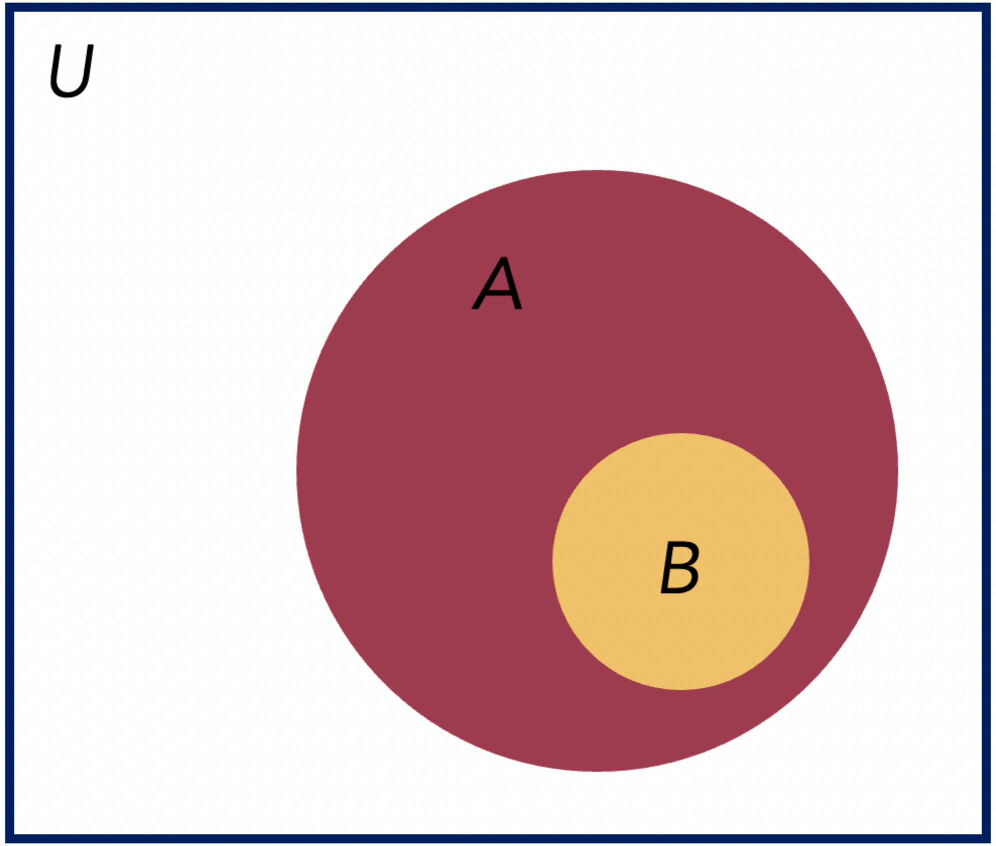
Questions

Part 2: Sets & Set Notation

Sets

- **Set**: Well-defined collection of objects; composed of **elements** or **members**
  - A = {Element 1, Element 2, Element 3,...Element i}
  - A = { x | x is a student at the University of Edinburgh}
- Elements in a set are represented with the following notation:
  - $x \in A$
    - x is an element of set A
  - $x \notin A$
    - x is not an element of set A
  - A = { x | x is an integer, $1 \leq x \leq 10$}
    - Set A includes elements *such that* x is an integer $\geq$ to 1 *and* $\leq$ to 10.

Sets



A = Set

U = Universal Set : All possible elements in a category of interest

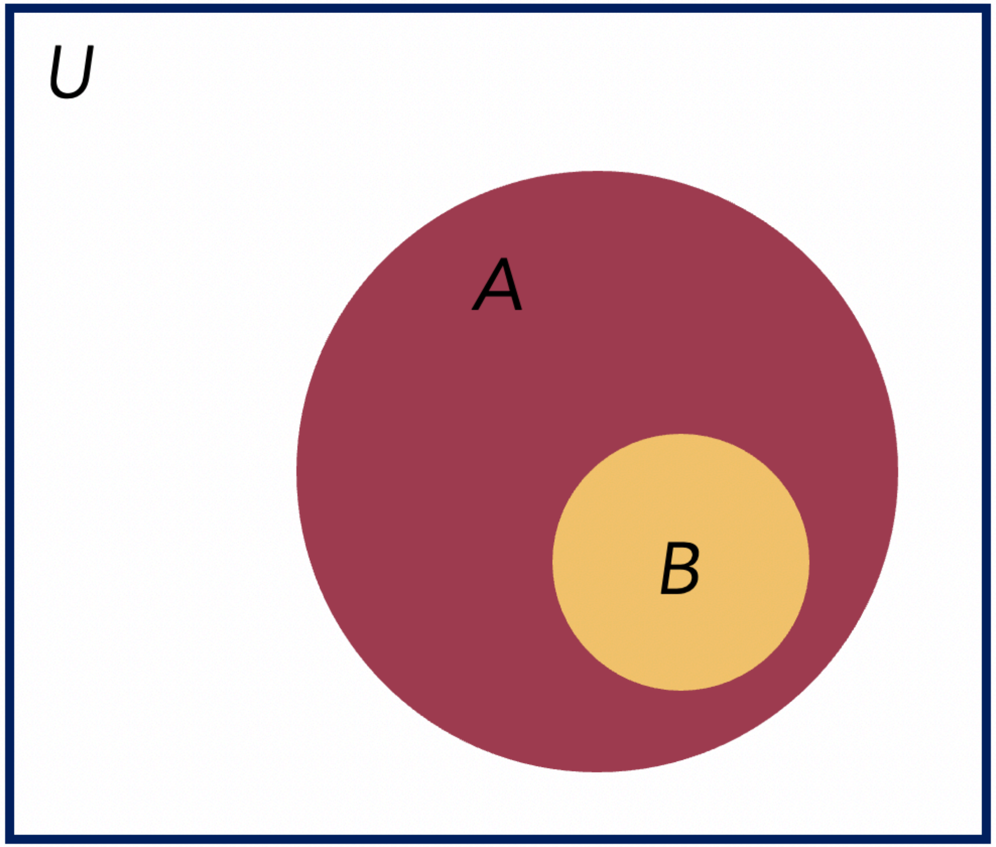
B = Subset

- If B is a subset of A :
  - All elements of B must also be in A
  - However, all elements in A do not have to exist in B (although they can)
  - E.g., $x \in B$ and $x \in A$. $y \in A$, but $y \notin B$

A^c = Complement of A

- $A^c = \{ x \mid x \in U, x \notin A\}$

Set Notation



$B \subseteq A$

- B is a subset of A

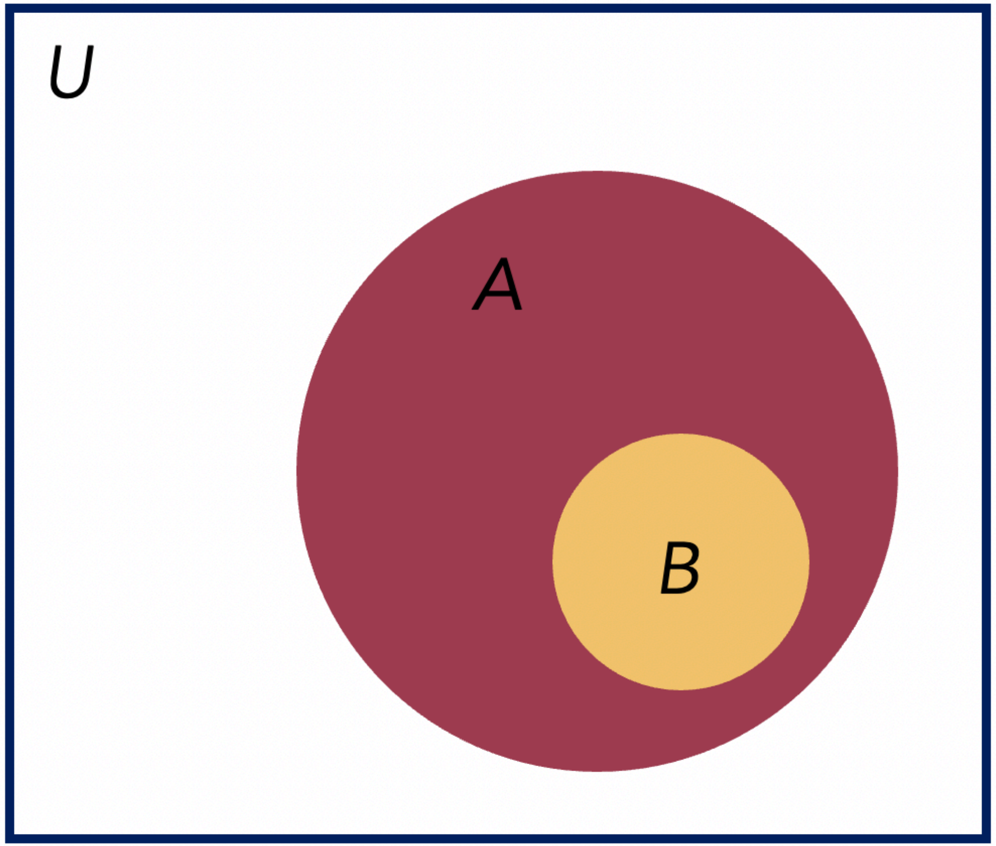
$B \subset A$

- B is a **proper** subset of A
- At least one element of A is **not** a member of B
- B is not identical to A

$A \not\subset B$

- A is not a subset of B

Set Example



U = All DapR1 students

A = DapR1 students who have a dog

B = DapR1 students who have a bulldog

A^c = DapR1 students who do not have a dog

$B \subseteq A$ because all bulldogs are dogs

$A \not\subset B$ because not all dogs are bulldogs

Set Operations

- There are also ways we can describe two distinct sets in terms of how they interact with each other.
    - **Union:** when an element is a member of either set A *or* set B (or both)
    - **Interaction:** when an element is a member of set A *and* set B
    - **Difference:** when an element is a member of set A *but not* set B, or vice versa
    - **Empty Set:** sets A and B are *mutually exclusive*; when A occurs, B cannot occur

Set Operations - Example Data

Imagine we have collected pet name data from 50 dog owners (Set A) and 50 cat owners (Set B):

```r
head(dogs, n = 10)
```

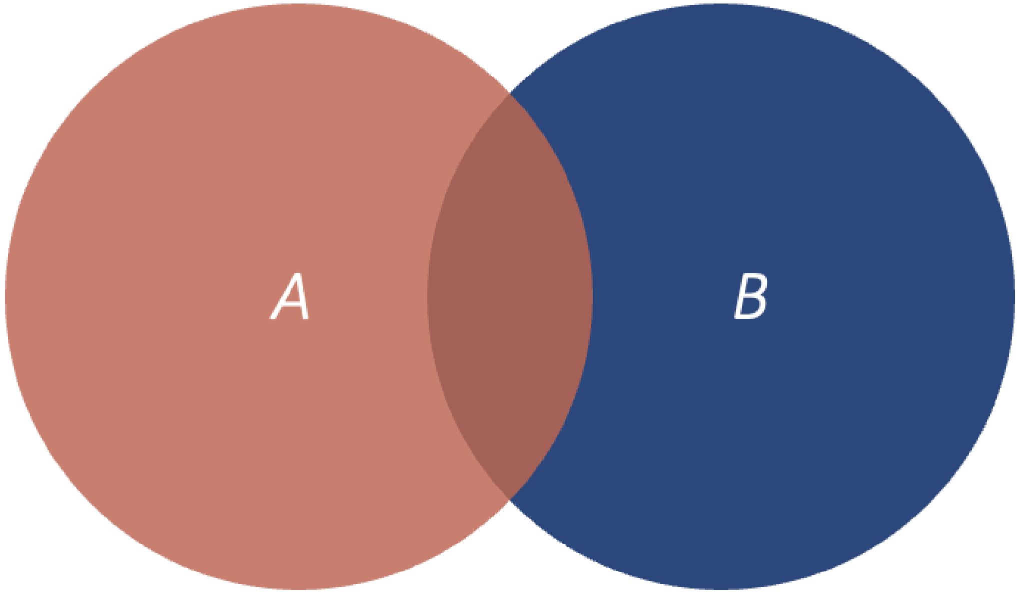
```
##          Name
## 1       Sooie
## 2     Charlie
## 3     Charlie
## 4       Moose
## 5         Jwl
## 6    Chiquita
## 7     Grommet
## 8      Metabo
## 9   Coco Rose
## 10     Caliie
```

```r
head(cats, n = 10)
```

```
##                    Name
## 1                 Mocha
## 2                Grumpy
## 3                  Luna
## 4   Sassafras (Sassy)
## 5                  Rasa
## 6                  Cleo
## 7                Keaton
## 8              Moonbeam
## 9                 Binks
## 10                Egypt
```

Set Operations - Union

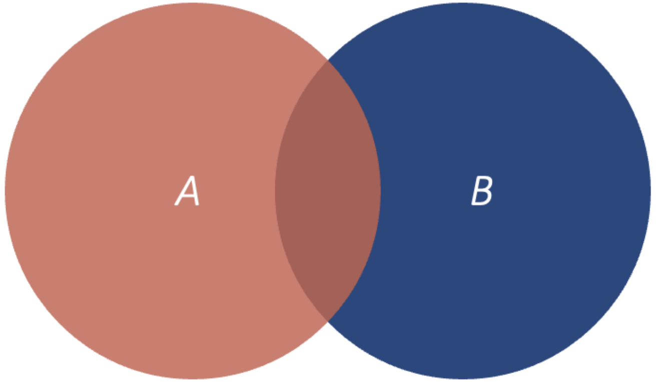
- A name is a member of either dogs **or** cats (or both)



$$A \bigcup B = \{\, x \mid x \in A \text{ or } x \in B \text{ or } x \in A \text{ and } B \,\}$$

Set Operations - Intersection

- A name is a member of both dogs **and** cats
- You can check the intersection of two sets in R using the `intersect` function



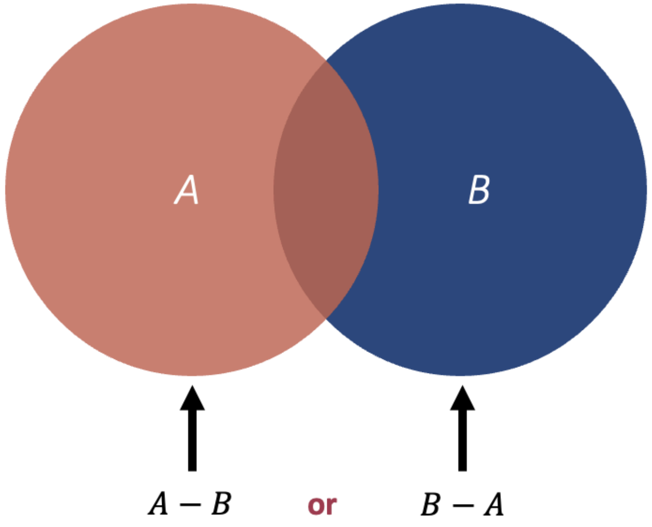
$A \bigcap B = \{ x | x \in A \textbf{ and } x \in B \}$

```
intersect(dogs, cats)
```

```
##      Name
## 1 Charlie
## 2    Milo
```

Set Operations - Difference

- A name used for a dog **but not** a cat, or vice versa
- You can check the difference between two sets in R using the `setdiff` function




```
head(setdiff(dogs, cats), 5)
```

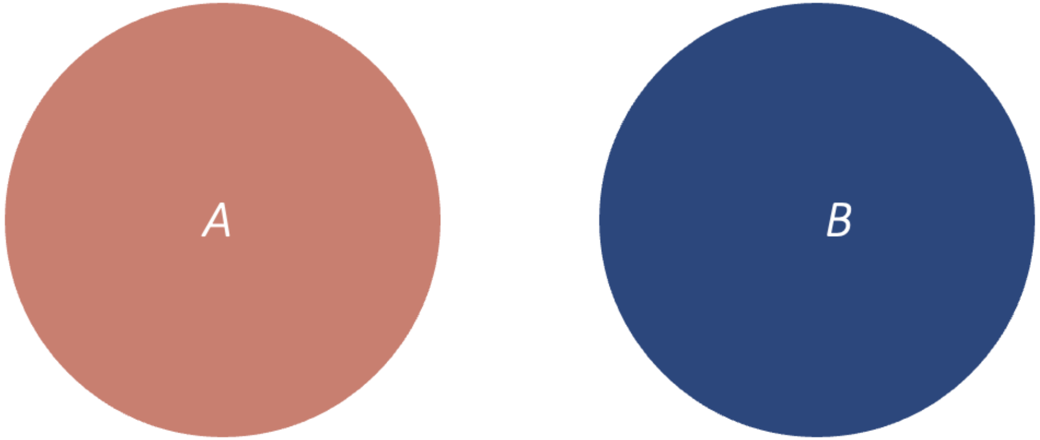
```
##        Name
## 1     Sooie
## 2     Moose
## 3       Jwl
## 4 Chiquita
## 5   Grommet
```

```
head(setdiff(cats, dogs), 5)
```

```
##                  Name
## 1               Mocha
## 2              Grumpy
## 3                Luna
## 4 Sassafras (Sassy)
## 5                Rasa
```

Set Operations - Empty Sets

- People who have a dog and people who don't own a pet are **mutually exclusive** groups; when one occurs, the other cannot



$$[A \bigcap B] = \emptyset$$

Questions

Part 3: Random Experiments

Random Experiments

- A procedure that meets certain criteria:
  - Can be repeated infinitely under identical conditions
  - Outcome depends on chance and can't be predicted in advance
- Are the following examples of random experiments?
  - Picking a card from a fair deck
  - Multiplying 8 and 6 on a calculator
  - Determining bus arrival times
- By conducting a random experiment, we can make inferences about the likelihood of each of its outcomes

Describing the Sample Space

- All possible outcomes of a random experiment are referred to as the **sample space** (S)
- Imagine a random experiment where you roll two six-sided dice, where the outcome is the sum of the roll.
- In this example, the total sample space, S , contains 36 elements

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Describing the Sample Space

- All possible outcomes of a random experiment are referred to as the *sample space* (S)
- Imagine a random experiment where you roll two six-sided dice, where the outcome is the sum of the roll.
- An **event**, A , is a subset of the outcomes from S
  - $A \subset S$
- A = At least one die landing on 4

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Describing the Sample Space

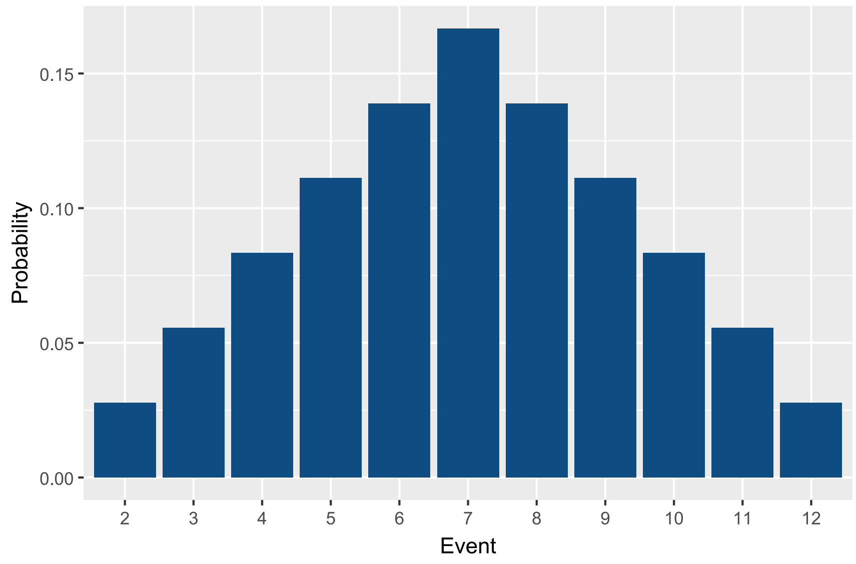
- All possible outcomes of a random experiment are referred to as the *sample space* (S)
- Imagine a random experiment where you roll two six-sided dice, where the outcome is the sum of the roll.
- A **simple event**, a , refers to a single element in a sample space
  - $a \in S$
- a = The dice summing to 12

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Visualising Probability

- A **probability distribution** is a mathematical function that describes the probability of each event within the sample space
- Plotting a probability distribution allows you to visualise the likelihood of all possible outcomes

Event	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Questions

Some of today's key takeaways

1. In statistics, we use probabilistic models to make inferences about our data
2. $P(x)$ is the proportion of times you would observe x if you took an infinite number of samples.
3. Random experiments refer to procedures that could be repeated infinitely and whose outcomes can't be predicted with certainty
4. We can use the results of random experiments to make inferences about the likelihood of each outcome

Next tasks

- Tomorrow, I'll present a live R session focused on sets and logical operators.
- Next week, we will discuss rules of probability
- This week:
  - Attend the live R session
  - Complete your lab
  - Check the reading list for recommended reading
  - Come to office hours
  - Weekly quiz
    - Opens Monday 09:00
    - Closes Sunday 17:00