Statistics Foundations: Understanding Probability and Distributions

INTRODUCING THE CONCEPT OF PROBABILITY

Dmitri Nesteruk
QUANTITATIVE ANALYST

@dnesteruk

Statistics

A branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.

How Statistics Works

Formulate a hypothesis

- "Smoking causes cancer"

Make observations

- Get data regarding smoking habits and medical history

Analyze data and make conclusions

- Accept or reject hypothesis
- Modify experiment, get more data, etc.

Course Overview

1. Introducing the Concept of Probability
2. Calculating the Conditional Probability of Events
3. Understanding Random Variables and Distributions
4. Introducing the Concept of Expectation
5. Looking at Some Special Statistical Distributions

Structure

First course in the Statistics Foundations series

Theory lectures

Live examples in R

Simulation/analysis comparison

Goal:

Understand the concept of *probability*; learn the basics of set theory and combinatorics.

Overview

Naïve Set Theory Primer

Experiments and Events

Probability

Counting Methods

Combinatorics

Probability of a Union of Events

Naïve Set Theory Primer

Set Theory

A branch of mathematics that deals with sets.

Naïve Set Theory = set theory explained in simple terms, without resorting to heavy math.

(Alternative: Axiomatic Set Theory)

Set

A well-defined collection of distinct objects.

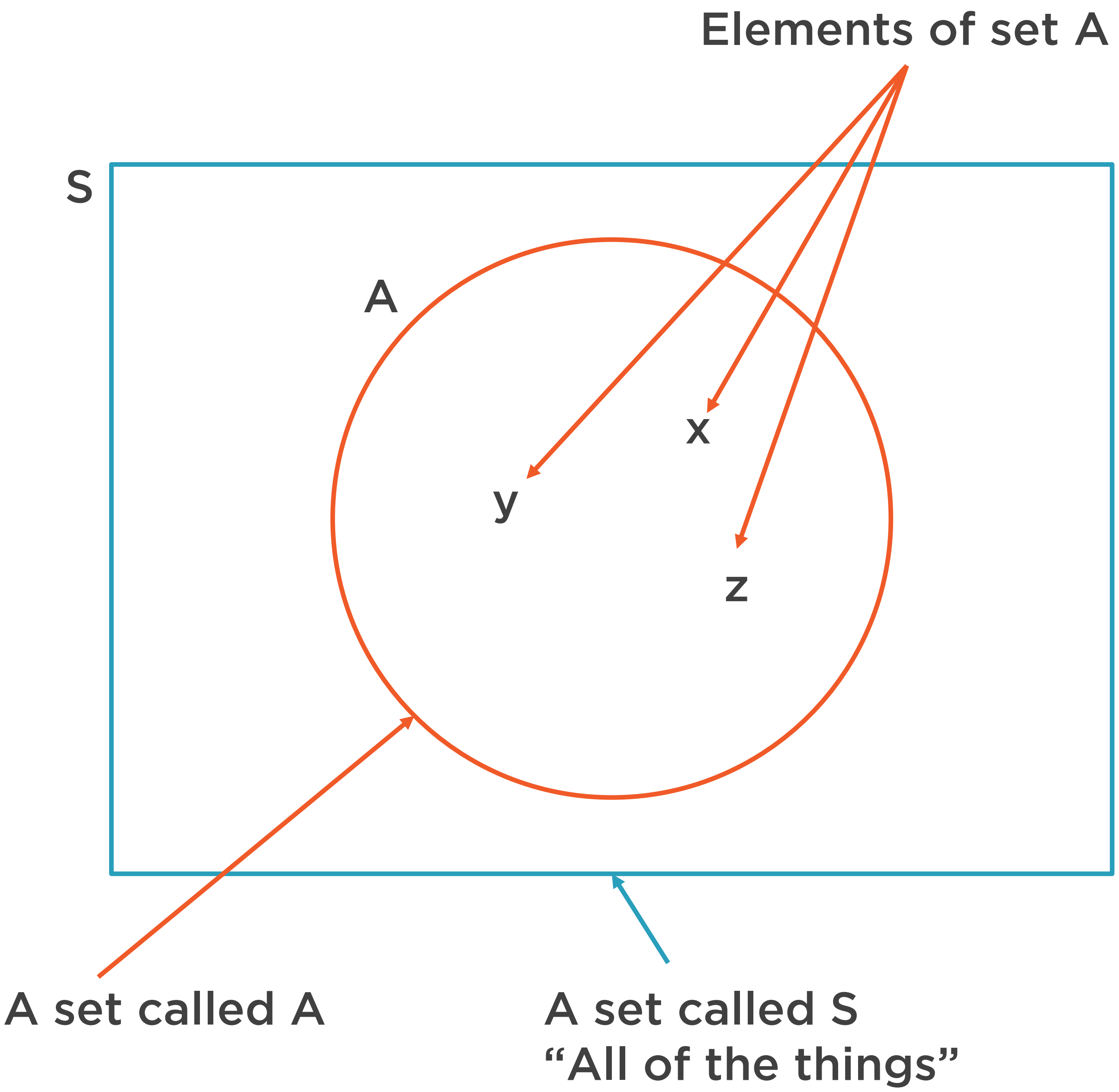
$A = \{0,1,2,3,4,5,6,7,8,9\}$ a set of digits

$B = \{a, b, c, \dots, z\}$ a set of letters

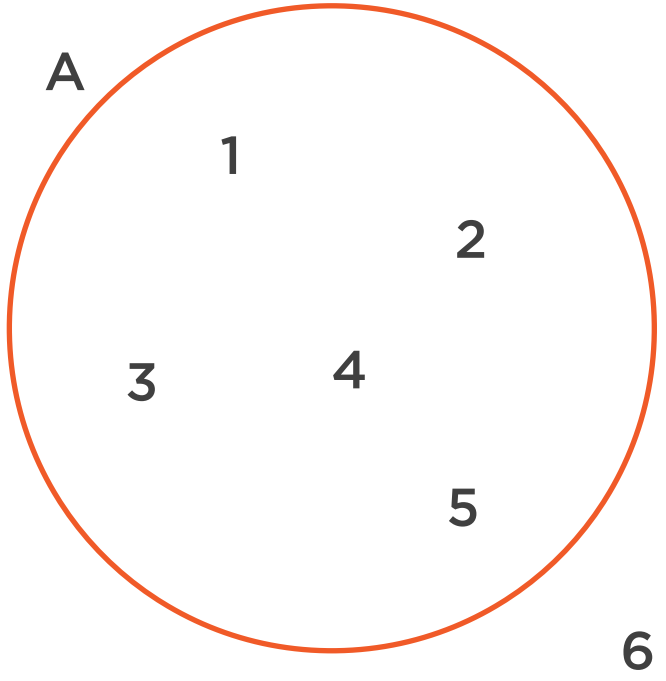
$S = \{\boxplus, \ominus, \circledast, \wedge\}$ a set of weird mathematical operators

Sets are often illustrated with Venn diagrams

Used for illustration in this course



Set Membership



Consider a set $A = \{1,2,3,4,5\}$

We can state that 1 is an element of A

$$1 \in A$$

And that 6 is *not* an element of A

$$6 \notin A$$

This notation is often used to specify a set of values

E.g., ‘for all even natural numbers’:

$$\forall x : \{x \in \mathbb{N} \mid x \text{ is even}\}$$

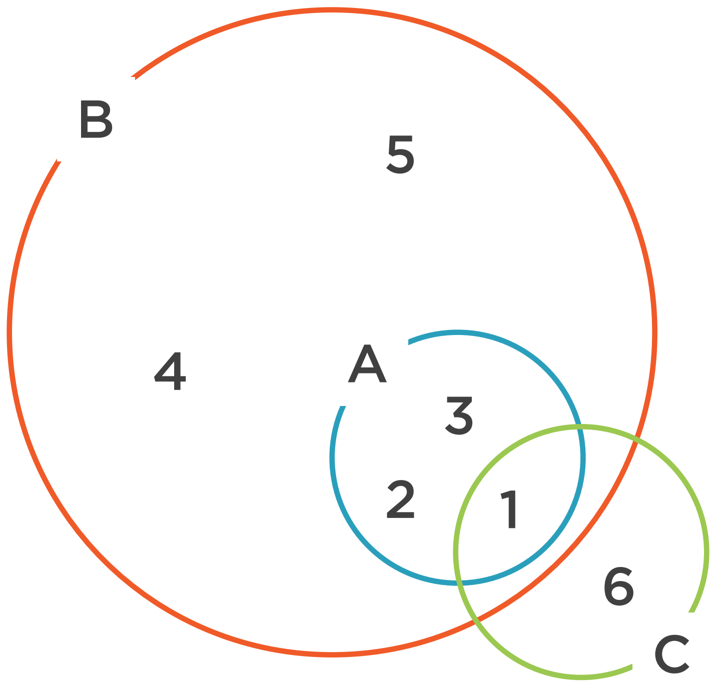
Empty Set

$\emptyset$

A set containing no elements

$\emptyset = \{\}$

For every element x, $x \notin \emptyset$



Subsets

Set A is a subset of set B if every element of A is in B

For example, given $A = \{1,2,3\}$, $B = \{1,2,3,4,5\}$, A is a subset of B:

$$A \subset B$$

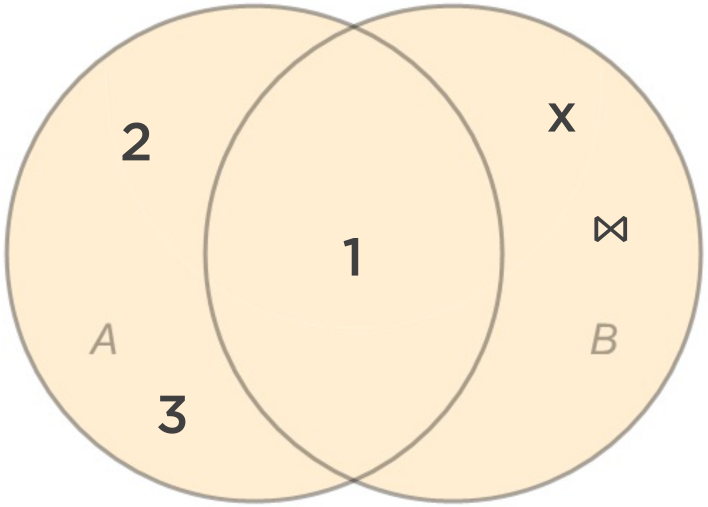
Similarly, we have a 'not a subset of' operator, so if $C = \{1,6\}$

$$C \not\subset B$$

For every set A

$$\emptyset \subset A$$
$$A \subset A$$

Set Union



A *union* of two sets A and B is a set containing all elements from A and B

E.g., if $A = \{1,2,3\}$ and $B = \{1, x, \bowtie\}$ then

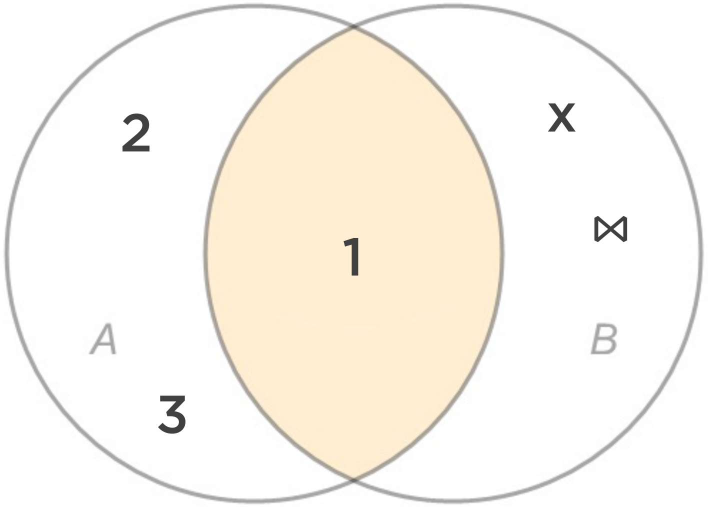
$$A \cup B = \{1,2,3, x, \bowtie\}$$

Notice that the common elements do not repeat

Also exists as a large operator

$$\bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3$$

Set Intersection



An *intersection* of two sets A and B is a set containing all that belong to both A and B

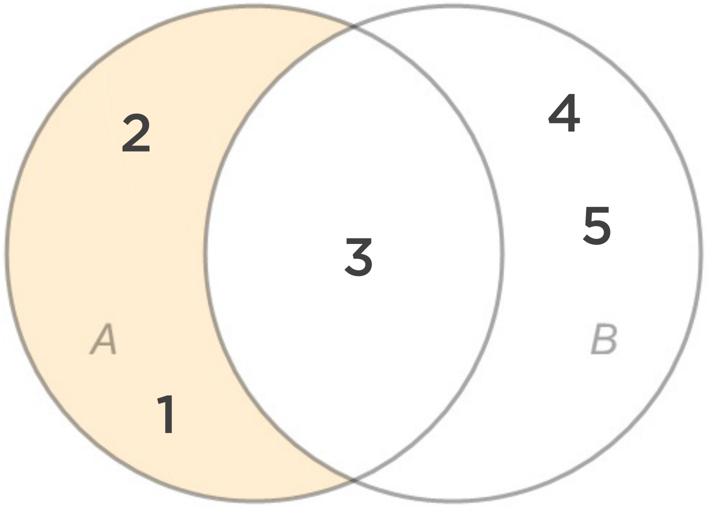
E.g., if $A = \{1,2,3\}$ and $B = \{1, x, \bowtie\}$ then

$$A \cap B = \{1\}$$

Also exists as large operator

$$\bigcap_{i=1}^{3} A_i = A_1 \cap A_2 \cap A_3$$

Set Difference



Given two sets A and B, the *set difference* is the set of items in A that are *not* in B

For example, if $A = \{1,2,3\}$ and $B = \{3,4,5\}$ then

$$A \setminus \mathrm{B} = \{1,2\}$$

Some obvious equalities

$$A \setminus A = \emptyset$$

$$\emptyset \setminus A = \emptyset$$

$$A \setminus \emptyset = A$$

Cardinality and Set Complement

Set *cardinality* = number of elements in a set

E.g., if $A = \{a, b, c\}, \#A = 3$

Set *complement* = set of elements not contained in A (but contained elsewhere)

E.g., if $E = \{2,4,6, \dots\}$ is a set of even natural numbers, then $E^c = \{1,3,5, \dots\}$ is the set of odd natural numbers

This can also be represented as

$$E^c = \mathbb{N} \setminus E$$

Die roll: if $A = \{1,2,5\}$, then, assuming $S = \{1,2,3,4,5,6\}$, $A^c = S \setminus A = \{3,4,6\}$

Some Set Laws

De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Distributive Properties

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

For every two sets A and B,

- $A \cap B$ and $A \cap B^c$ are disjoint
- $A = (A \cap B) \cup (A \cap B^c)$

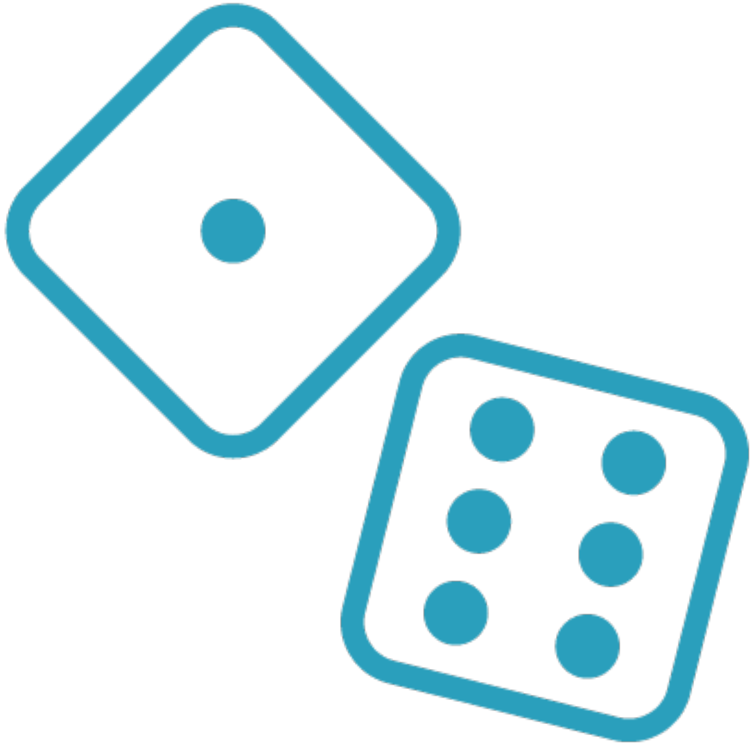
A lot more laws and set operations; consult a textbook

Demo

Set Theory in R

Experiments and Events

Events



Consider an experiment consisting of a single balanced 6-sided die roll

The value rolled is a *random variable*

Player rolled a 6

- This is a *simple event*: it cannot be decomposed

Player rolled an even number

- Player rolled a 2, 4 or 6
- This is a *complex event*: it can be decomposed into simpler events

Sample Space

The set of all possible values of a random variable is called the *sample space*.

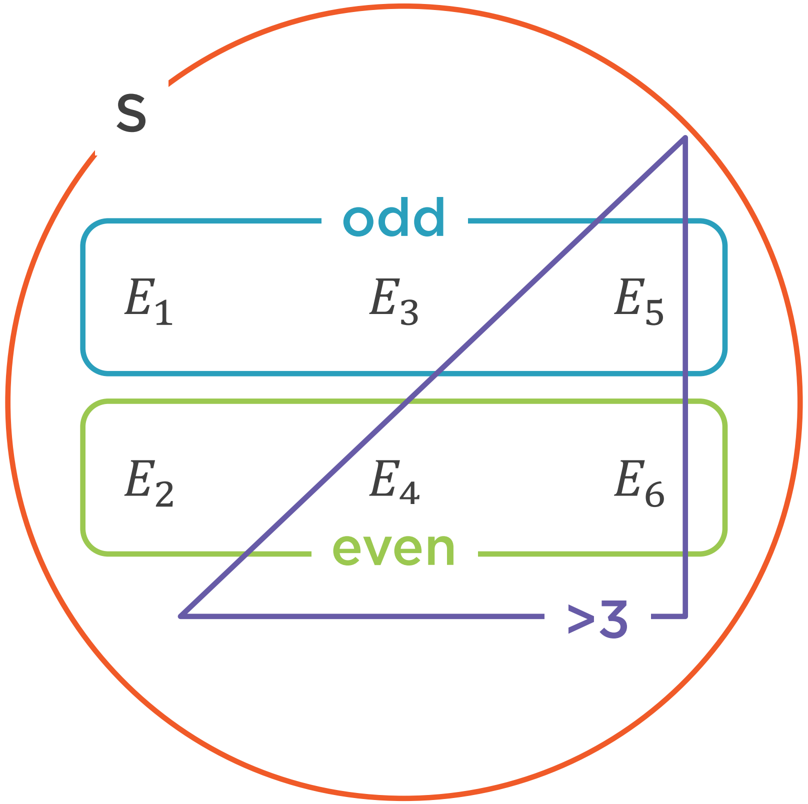
The sample space for a die roll is

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

where E_n is an event when n is rolled.

Each element of the sample space is a *sample point*.

Set Operations on Events



Suppose a die roll is

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

An even roll

$$S_{\text{even}} = \{E_2, E_4, E_6\}$$

An roll that is greater than 3

$$S_{>3} = \{E_4, E_5, E_6\}$$

A roll that is even *<u>or</u>* greater than 3

$$S_{\text{even}} \cup S_{>3} = \{E_2, E_4, E_5, E_6\}$$

A roll that is even *<u>and</u>* greater than 3

$$S_{\text{even}} \cap S_{>3} = \{E_4, E_6\}$$

A roll that is odd *<u>and</u>* greater than 5

$$S_{\text{odd}} \cap S_{>5} = \emptyset$$

Independence

If events do not influence one another, they are called *independent*

- Successive die rolls or coin flips

If subsequent events influence one another, they are *dependent*

- Colored balls being pulled out of a hat without replacement
- The set of possible balls gets reduced as you pull them out

Probability

Probability

How likely an event is to occur.

Probability

Probability of an event is a number that is

- A value between 0 and 1 inclusive
- Value of 0 corresponds to “unlikely”, “almost never”
- Value of 1 corresponds to “almost surely”, abbreviated a.s.

Typically recorded as $P(\text{event}) = \text{value}$

Sometimes also expressed as percentage (0% to 100% respectively)

- There’s a 50% chance that → $P = 0.5 = \frac{1}{2}$
- I am 90% certain that → $P = 0.9 = \frac{9}{10}$
- Used in e.g., Excel

Rules of Probability

Probability of event A is greater than or equal to zero

$$P(A) \geq 0$$

Probability of sample space is one

$$P(S) = 1$$

If $A_1, A_2, A_3, \ldots$ are a sequence of mutually exclusive events ($A_i \cap A_j = \emptyset$ if $i \neq j$), then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \bigcup A_i = \sum P(A_i)$$

Probability Examples

Consider a coin flip $S = \{E_H, E_T\}$

$P(S) = 1$ by definition

$P(E_H) = P(E_T)$ as both events are equally likely

Solving $\begin{cases} P(E_H) + P(E_T) = 1 \\ P(E_H) - P(E_T) = 0 \end{cases}$ gives us

$P(E_H) = P(E_T) = \frac{1}{2}$

$P(\text{head and tail}) = 0$ (coin cannot on both head and tail at the same time)

$P(\text{head or tail}) = P(S) = 1$ a.s.
(coin will definitely land on head or tail)

Demo

Basic Probability Simulation

Discrete vs Continuous

Discrete = finite set of unique values

- Coin toss, die roll, number of cars in a household
- $P(E) = \frac{1}{N}$ given N possible events, assuming all events equally likely

Continuous = infinite set of values

- Person's height, amount of rainfall in a day
- $P(\text{you are 1.77m tall}) = 0$ unless you round people's height
- Instead, better to measure intervals, e.g., $P(1.77 < h < 1.78)$

Counting Methods

Counting Sample Points

Consider a coin flip $S = \{E_H, E_T\}$

Flip a coin three times (or flip 3 coins once)

- Sampling *with* replacement

How many different arrangements are possible?

Direct approach: list all the arrangements and count them

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

How many arrangements begin or end with a head? $\#(S_{H??} \cup S_{??H})$

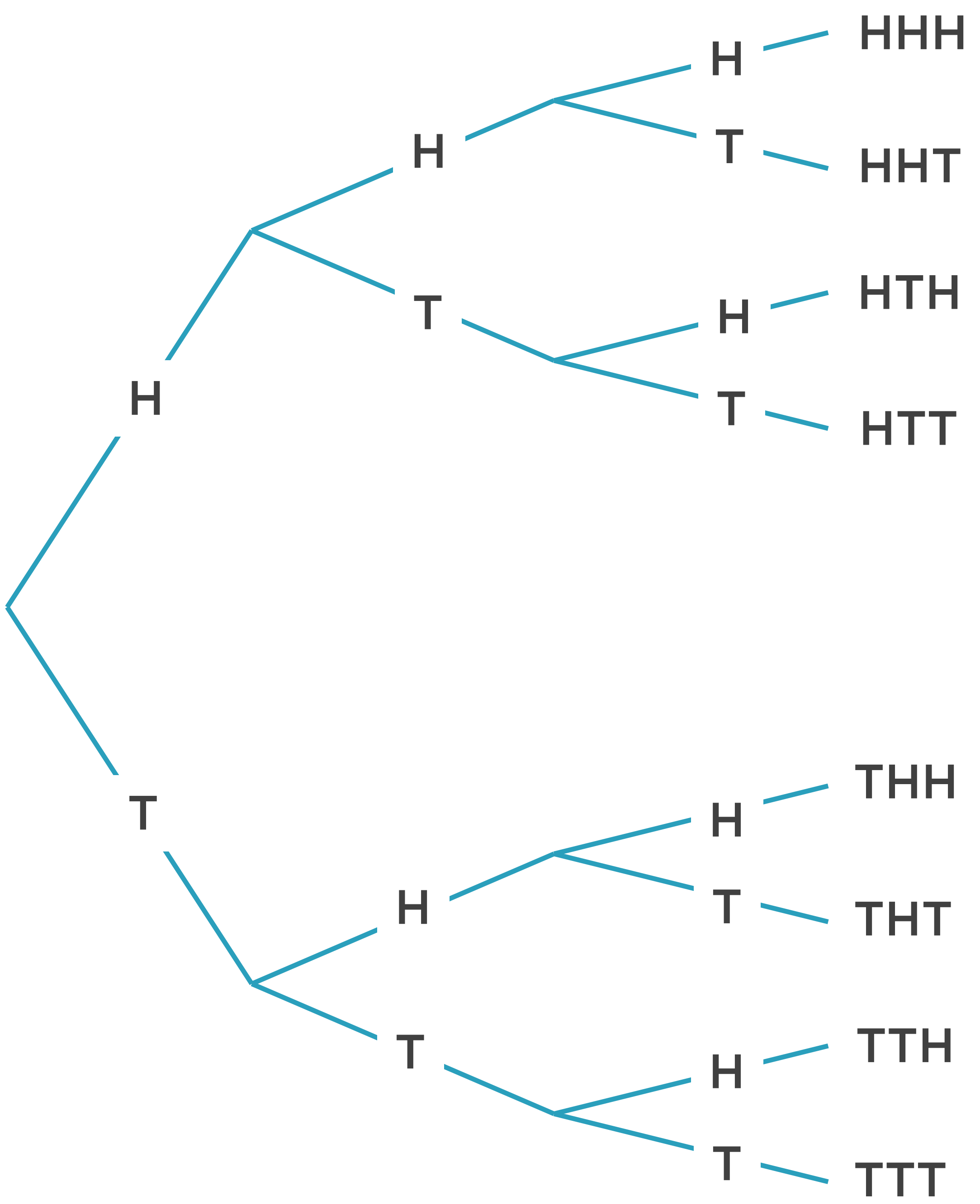
Again, count them: 6

We can also visualize the sample space as a tree

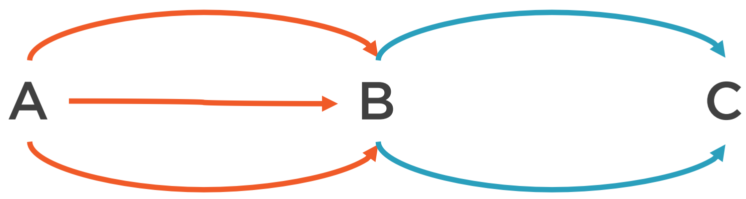
Each branch represents a possible event

Final branches represent all possible outcomes

Clearly, given N tosses, total # of outcomes = 2^N



Multiplication Rule



We need to travel from A to C through B

Three ways of getting from A to B

- $AB = \{AB_1, AB_2, AB_3\}$

Two ways of getting from B to C

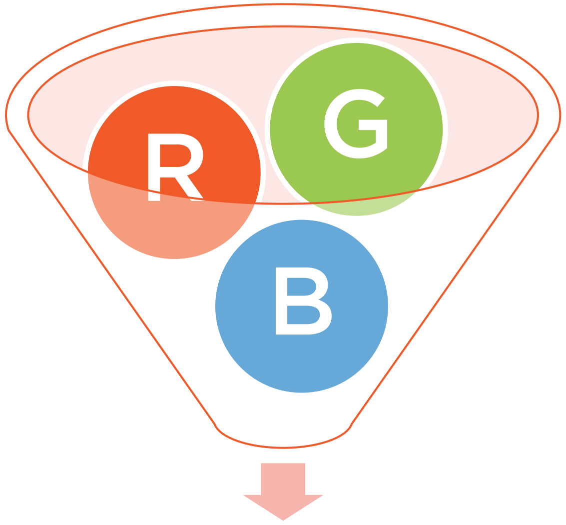
- $BC = \{BC_1, BC_2\}$

How many ways of getting from A to C?

	AB_1	AB_2	AB_3
BC_1	(AB_1, BC_1)	(AB_2, BC_1)	(AB_3, BC_1)
BC_2	(AB_1, BC_2)	(AB_2, BC_2)	(AB_3, BC_2)

Clearly, number of options is

$$\#AB \cdot \#BC = 3 \cdot 2 = 6$$



A bag has three balls: red, green, blue

You draw each of the balls in turn *without replacement* (you keep the ball)

How many possible draws are there?

First draw: three possible options

Second draw: one ball already taken, two possible options

Last draw: only one option

Thus, total number of possible draws is:

$$3 \times 2 \times 1 = 6$$

They are: RGB, RBG, GRB, GBR, BRG, BGR

Permutations

Given a set of elements, all distinct arrangements of these elements are called the *permutations* of a set

The number of ways you can arrange N elements out of N possibilities is…

$P_{n,n} = N \times (N-1) \times N - 2 \times \cdots \times 2 \times 1 = N!$ **permutations**

N! = "N factorial" = a product of all numbers from N down to 1

- Assumes N is a non-negative integer
- By definition, $0! = 1$

General Formula for Permutations

Suppose there are $n = 5$ balls (RGBWO) and you draw $k = 3$ without replacement

How many possible arrangements are there?

5 possibilities on first draw, 4 on the second, 3 on the third, so $5{\times}4{\times}3 = 60$

But how to express it in terms of n and k?

$$5{\times}4{\times}3 = \frac{5{\times}4{\times}3{\times}2{\times}1}{2{\times}1} = \frac{5!}{2!} = \frac{n!}{(n-k)!}$$

Thus, the number of permutations of k elements out of n is $P_{n,k} = \frac{n!}{(n-k)!}$

Permutation Examples

Number of unique 5-letter words that can be arranged from letters $\{a, b, c, d, e\}$

$$5^5 = 3125$$

If each letter appears exactly once

$$5! = 120$$

3-letter words, each letter appears once

$$P_{5,3} = \frac{5!}{(5-3)!} = 60$$

Birthday Problem

Given a room with k people, what is the probability that *at least* two people have the same birthday?

- Ignore leap years, seasonal variations

Obvious: if $k > 365$, $P = 1$, so assume $k \leq 365$

Number of arrangements when each birthday is different is $P_{365,k}$

Total number of arrangements $= 365^k$

P(no two people have same b/day) $= \frac{P_{365,k}}{365^k}$

∴P(two+ people have same b/day)

$$= 1 - \frac{P_{365,k}}{365^k}$$

Demo

Birthday Problem

Combinatorics

Combinatorics

Combinatorics is a branch of mathematics concerning the study of finite or countable discrete structures.

Combinatorial methods

Given a bag of balls $B = \{R, G, B, W, O\}$

If we pull out 2 balls after replacement, the number of arrangements is

$$P_{5,2} = \frac{5!}{3!} = 20$$

What is we don’t care about order, i.e., we consider $\{R, B\}$ and $\{B, R\}$ to be the same pick?

How can we calculate the number of unique picks now?

Problem: find the number of subsets of a set (remember, order doesn’t matter!)

Combinatorial methods

Given a bag of balls $B = \{R, G, B, W, O\}$

The total number of picks (including duplicate sets) is $P_{5,2}$

- And each of the picks has 2! possible arrangements

So we reduce the total number of picks by the number of arrangements in each pick

$$C_{n,k} = \frac{P_{n,k}}{k!} = \frac{n!}{k!\,(n-k)!}$$

In our case, $C_{5,2} = \frac{5!}{2!(5-2)!} = 10$

Here they are:

$\{RG, RB, RW, RO, GB, GW, GO, BW, BO, WO\}$

Binomial Coefficients

We also denote $C_{n,k}$ as $\binom{n}{k}$

$\binom{n}{k}$ is called a *binomial coefficient* because, according to the Binomial Theorem,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

$\forall n, \binom{n}{0} = \binom{n}{n} = 1$

Also, $\binom{n}{k} = \binom{n}{n-k}$

Binomial Coefficient Examples

You toss a coin 10 times and record the result. What is the probability of getting exactly 4 heads?

There are a total of 2^{10} arrangements

Each arrangement is a choice as to where put the 4 heads among 10 tosses; $\binom{10}{4}$ arrangements

$$\therefore P(4\text{heads}) = \frac{\binom{10}{4}}{2^{10}} = 0.205$$

Number of sequences of 4 heads *or less*

$$P(\leq 4\text{heads}) = \frac{\sum_{i=0}^{4}\binom{10}{i}}{2^{10}} = 0.377$$

Binomial Coefficient Examples

A class has 15 girls and 30 boys

Pick 10 children at random. What's the probability you'll pick exactly 3 girls?

Number of ways of picking 3 girls from 15 girls is $\binom{15}{3}$; number of ways of picking 7 boys from 30 is $\binom{30}{7}$

Overall number of combinations is $\binom{45}{10}$

$$\therefore P(3 \text{ girls}) = \frac{\binom{15}{3}\binom{30}{7}}{\binom{45}{10}} = 0.29$$

Multinomial Coefficients

10 students need to form 3 groups consisting of 4, 3 and 3 members respectively

How many ways can students be assigned to these groups?

First group: choose 4 students out of 10, $\binom{10}{4}$ arrangements

We are left with 6 students; number of ways to split them is $\binom{6}{3}$

$$\binom{10}{4}\binom{6}{3} = \frac{10!}{4!6!} \cdot \frac{6!}{3!3!} = \frac{10!}{4!3!3!} = 4200$$

Multinomial Coefficients

In general, number of arrangements of n elements into k groups of size $n_i = \{n_1, n_2, \dots\}$ is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots = \frac{n!}{n_1!\,n_2!\cdots n_k!}$$

This is the *multinomial coefficient*, written as

$$\binom{n}{n_1, n_2, \dots, n_k}$$

Just like with binomial coefficients,

$$(x_1 + \cdots + x_k)^n = \sum \binom{n}{n_1, n_2, \dots, n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$$

Multinomial Coefficient Examples

Picking 4,3,3 students out of 10

$$\binom{10}{4,3,3} = \frac{10!}{4!\,3!\,3!} = 4200$$

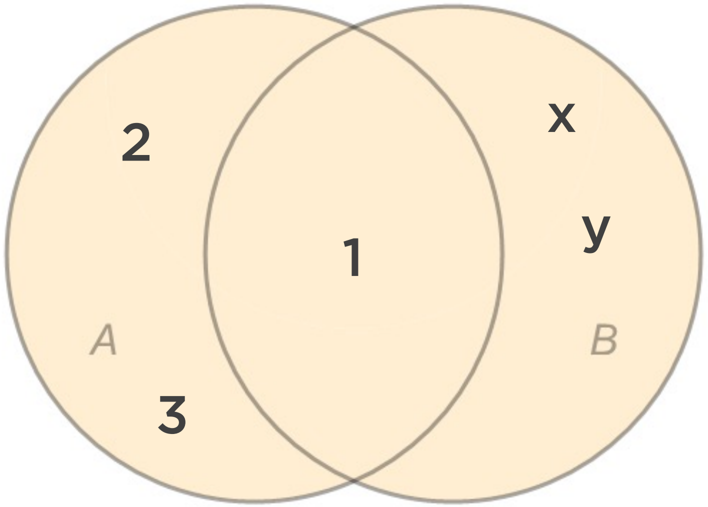
Number of ways to arrange 3 a's, 4 b's and 5 c's is

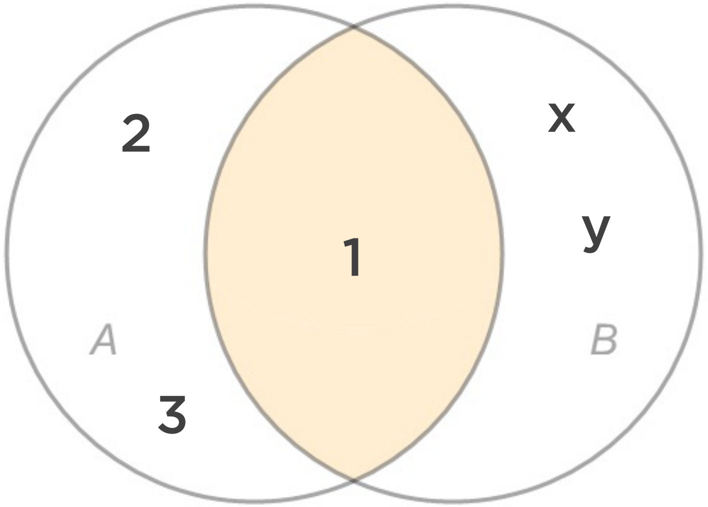
$$\binom{12}{3,4,5} = \frac{12!}{3!\,4!\,5!} = 27{,}720$$

In the expansion of $(x+y+z)^3$, the coefficient of the term $x^2z = x^2y^0z^1$ is

$$\binom{3}{2,0,1} = \frac{3!}{2!\,0!\,1!} = 3$$

Probability of a Union of Events





We also know that for a set of disjoint events A_i

$$P\left(\bigcup A_i\right) = \sum P(A_i)$$

For every two (not necessarily disjoint) events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This can be extended to 3 or more events, e.g.,

$$\begin{aligned} &P(A \cup B \cup C) \\ &= P(A) + P(B) + P(C) \\ &- \{P(A \cap B) + P(B \cap C) + P(A \cap C)\} \\ &+ P(A \cap B \cap C) \end{aligned}$$

Union of Events Example

Consider a cohort of 200 students

50 students take programming (P), 100 students take electronics(E), 150 students take maths (M)

30 students take programming+electronics, 45 students take electronics+maths, 25 students take electronics+programming

15 students take all 3 classes; some students take no classes from the list

What is the probability that a student takes at least one class?

Union of Events Example

$$P(P) = \frac{50}{200}, P(E) = \frac{100}{200}, P(M) = \frac{75}{200}$$

$$P(P \cap E) = \frac{30}{200}, P(E \cap M) = \frac{45}{200}, P(E \cap P) = \frac{25}{200}$$

$$P(P \cap E \cap M) = \frac{15}{200}$$

$$\therefore P(P \cup E \cup M)$$
$$= \frac{50}{200} + \frac{100}{200} + \frac{75}{200} - \left\{\frac{30}{200} + \frac{45}{200} + \frac{25}{200}\right\} + \frac{15}{200}$$
$$= \frac{140}{200} = 0.7$$

Summary

To study phenomena, we perform experiments and record our observations

The sample space describes all possible outcomes; these can be modeled with sets and measured using counting or combinatorics methods

Probability: value in range 0 to 1 inclusive; described the likelihood of an event occurring

If events are mutually independent, P(union of events) = sum of their individual probabilities