MAT3026 - Probability and Statistics

Lecture 1 – Introduction, Counting and Probability

- Introduction to probability and statistics

Walpole Chapter 1

See also the Appendix covering:

- *Sample Space* *Walpole Section 2.1*
- *Events* *Walpole Section 2.2*

Introduction to Probability and Statistics

Probability

Although physics tells us that the (macroscopic)universe is mechanical and deterministic, processes are generally very complex which leads to the notion of *randomness.*

We say *random* because the process is too complex or too difficult to calculate in a deterministic way. To predict the behaviour of such systems we need to assign *probabilities* to *observed outcomes* and use those probabilities to describe(model) the process, and to predict future outcomes.

In this way, we can consider questions such as:

- How are the outcomes of this process distributed? What is the average outcome?
- Are these two outcomes independent, or are they connected by a common cause?
- What is the probability that this system of components will work?

Statistics

In the last four lectures of the course, we will build statistical tools that allow us to make meaningful inferences about population parameters based upon sampling of populations.

An important principle here is quantifying uncertainty;
"A measurement is meaningless without an estimation of its uncertainty!"

With statistical tools and procedures, we can consider questions such as:

- How accurate and precise is this process?
- Is this observation consistent with our assumption about the process?
- What is the average outcome of a process, and how confident are we about that?

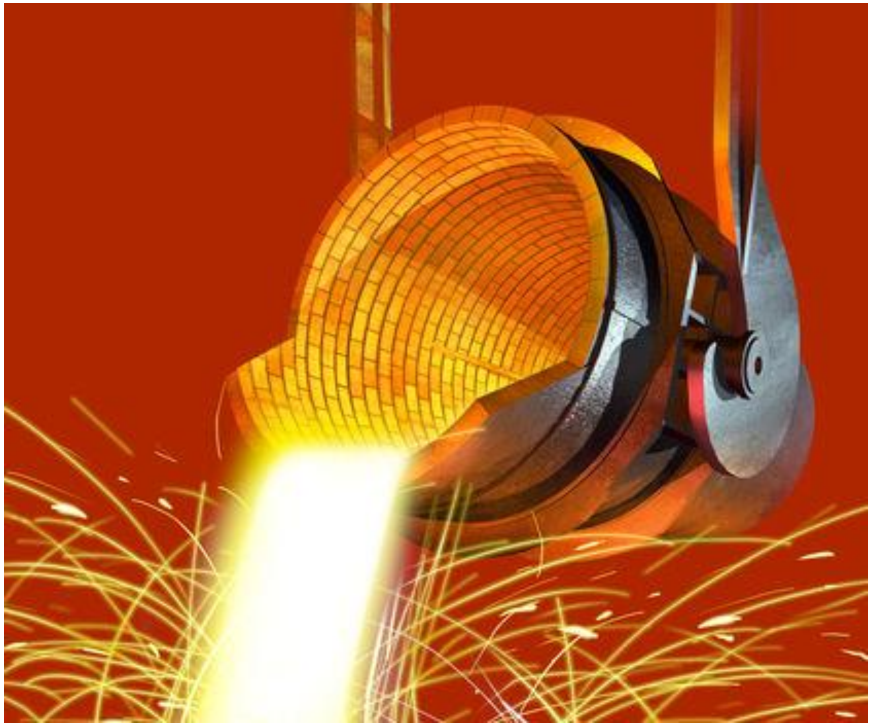
A knowledge of probability and statistics allow us to understand data and make informed engineering decisions.

Variability in Processes

Consider a manufacturing process where a machine smelts iron ingots.

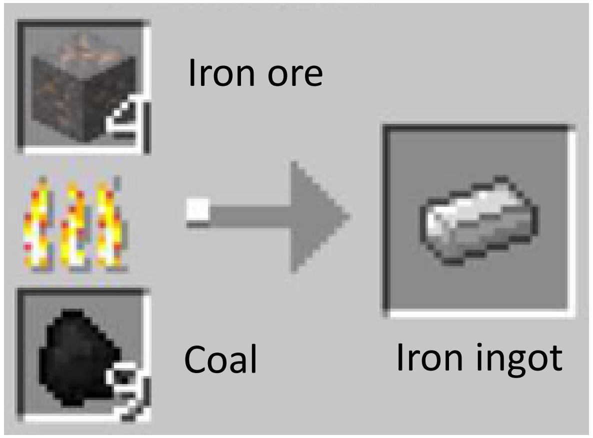
The goal is to create a population of identical ingots, but in reality the interaction of complex processes results in a small variation in the size and constituents (and therefore mass) of the ingots.

http://www.pollutionissues.com/Re-Sy/Smelting.html

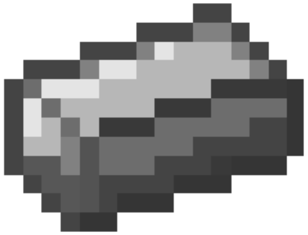


In the minecraft game, all iron ingots are identical!

Smelting (minecraft)



http://minecraft.gamepedia.com/Iron_Ingot



However, in reality, there will be some variation from one ingot to the next.

www.haotianmetal.net

100,000 ingots.

The mass of the ingots varies slightly from item to item.

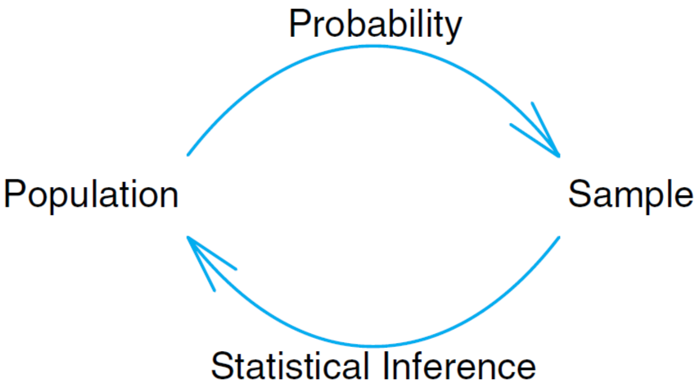
Sampling populations

Random sampling of the population.

www.haotianmetal.net

100,000 ingots.

The mass of the ingots varies slightly from item to item.



Probability

Population

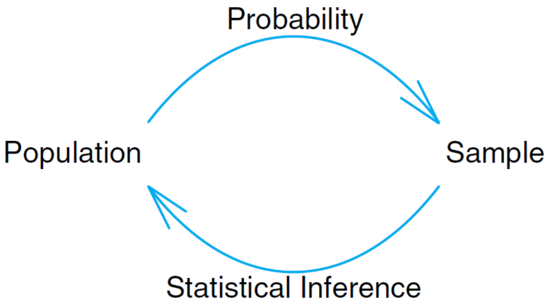
Sample

Statistical Inference

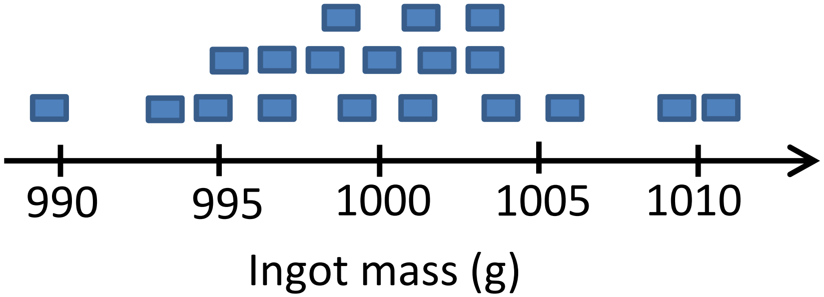
A sample of 19 ingots.

From this sample, what can we say about the properties of the whole population?

A sample of 19 ingots.



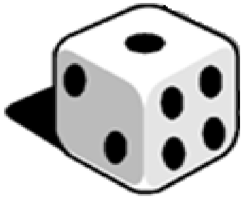
From this sample, what can we say about the property of the whole population?



Roughly speaking, we can expect any randomly-selected ingot to be in the range 1000 ± 10 g.

More usefully, we can say ***most*** of the ingots are in the range 1000 ± 5 g.

We will see later that *statistics* will provide us with a **formal** way to state this, and to infer other properties of the population.



“Random” Processes

A basic example of a complex *random* system: throwing a die:

The probability of obtaining a ⚃ is 1/6;

We believe this to be correct because there are six possible outcomes, all are equally likely (for a fair die), and only one outcome is a ⚃.

Formally: The probability of obtaining an outcome of interest =

$$\frac{\text{“number of outcomes of interest”}}{\text{“number of all possible outcomes”}}$$

and so:

$$P(\text{⚃}) = \frac{n\{\text{⚃}\}}{n\{\text{⚀ ⚁ ⚂ ⚃ ⚄ ⚅}\}} = \frac{1}{6} \approx 0.167$$

If you are not familiar with sets, events, sample spaces then see the notes in the Appendix.

https://en.wikipedia.org/wiki/Probability

In general, we will talk about "*the probability of an event*"

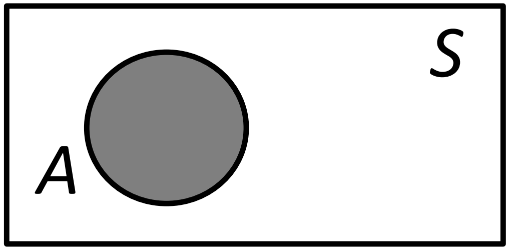
Given a sample space *S* (the Universal set of all possible outcomes) and event *A*; the probability that the outcome of the random process will be contained within event *A* is given by:

$$P(A) = \frac{n(A)}{n(S)}$$

← number of outcomes in collection *A*

← number of all possible outcomes

assuming every sample point is equally likely!



The probability of an outcome being an element in *S* is 1 (a certainty):

$P(S) = 1$, and $P(\phi) = 0$

Therefore, any probability must be contained within the range: $0 \leq P(A) \leq 1$

That's the theory, but how can we measure this experimentally?

Probability (experimental interpretation)

If a random process is repeated n times (i.e. there are n trials) and event *A* occurs *n(A)* times then the probability of event *A* is defined as a *relative frequency* below:

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

number of occurrences of event *A*

number of trials

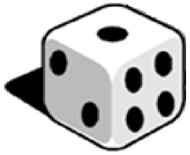
$P(A)$ approaches the true probability as n approaches infinity.

A basic property of probability

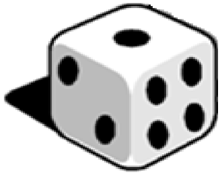
$$0 \le P(A) \le 1$$

$n(A)$=0
no outcomes
are event A

$n(A)$=n
all outcomes
are event A



To determine the probability of obtaining a ⚃ on the throw of a die; n throws would represent n trials and event *A* = {⚃}.



The experiment

Throw a die n times and count the number of times a ⚃ occurs.

n	n(⚃)	P(⚃) = n(⚃)/n	rel. error	P (≈ ⅙)
1	0	0.0	-100%	*Nan*
10	2	0.2	+20%	1/5.000 000
100	18	0.18	+8%	1/5.555 556
1000	161	0.161	-3%	1/6.211 180
10^4	1673	0.1673	+0.38%	1/5.977 286
10^5	16788	0.16788	+0.73%	1/5.956 635
10^6	166835	0.166835	+0.10%	1/5.993 946
10^7	1665173	0.1665173	-0.09%	1/6.005 382
10^8	16670178	0.16670178	+0.021%	1/5.998 736
10^9	166667244	0.166667244	+0.0003%	**1/5.999 979**

we expect

$$P = 1/6$$

$$\text{error} = P - 1/6$$

relative error

$$= \frac{P - 1/6}{1/6}$$

$$error \propto \frac{1}{\sqrt{n}}$$

COMPUTER PROGRAMMING IS
NOT EXAMINED IN THIS COURSE

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

Of course we cannot throw a die 10^9 times by hand!
The experiment in this case is a computer simulation:

```
//
// Computer simulation of the statistics of throwing a die.
//
#include <iostream>
#include <cstdlib>

int main() {

  const int n = 1000000000;  // One billion trials
  int m = 0;                 // This var will count the number of "4"s

  for (int i=0; i<n; i++) {  // Perform n trials
    int a = rand() % 6 + 1;  // Random outcomes 1,2,3,4,5,6.
    if (a==4) m++;           // Count the number of "4"s.
  }

  std::cout << n << " " << m << " " << double(m)/n << std::endl;
}
```

C++

```
1000000000 166667244  0.16666724
```

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

$\Rightarrow P(\text{“4”}) = 1/5.999979$

$$P(A) = \frac{n(A)}{n(S)}$$

Now consider throwing **two** dice

The set of all possible outcomes is:

```
{11,12,13,14,15,16,
 21,22,23,24,25,26,
 31,32,33,34,35,36,
 41,42,43,44,45,46,
 51,52,53,54,55,56,
 61,62,63,64,65,66}
```

i.e. $6^2 = 36$ permutations (order is counted).

What is the probability of obtaining **two 4's**?

Solution: there is one such outcome in 36 and so *P*("44") = 1/36.

What is the probability of obtaining **a 3 and a 4 (in any order)**?

Solution: there are two such outcomes in 36 and so *P*("34" $\cup$ "43") = 2/36.

$$P(A) = \frac{n(A)}{n(S)}$$

Now consider throwing **three** dice

There are
$6^3 = 216$ permutations

What is the probability of obtaining exactly two 4’s? (⚃ ⚃) in any order.

Solution: there are 15 such outcomes out of 216, so the probability = 15/216.

The set of all possible outcomes is:

{111 112 113 114,115,116,
121,122,123,124,125,126,
131,132,133,134,135,136,
141,142,143,**144**,145,146,
151,152,153,154,155,156,
161,162,163,164,165,166,
211,212,213,214,215,216,
221,222,223,224,225,226,
231,232,233,234,235,236,
241,242,243,**244**,245,246,
251,252,253,254,255,256,
261,262,263,264,265,266,
311,312,313,314,315,316,
321,322,323,324,325,326,
331,332,333,334,335,336,
341,342,343,**344**,345,346,
351,352,353,354,355,356,
361,362,363,364,365,366,
411,412,413,**414**,415,416,
421,422,423,**424**,425,426,
431,432,433,**434**,435,436,
441,**442**,**443**,444,**445**,**446,**
451,452,453,**454**,455,456,
461,462,463,**464**,465,466,
511,512,513,514,515,516,
521,522,523,524,525,526,
531,532,533,534,535,536,
541,542,543,**544**,545,546,
551,552,553,554,555,556,
561,562,563,564,565,566,
611,612,613,614,615,616,
621,622,623,624,625,626,
631,632,633,634,635,636,
641,642,643,**644**,645,646,
651,652,653,654,655,656,
661,662,663,664,665,666}

$P(X=0) = 125/216$, $P(X=1) = 75/216$, **$P(X=2) = 15/216$**, $P(X=3) = 1/216$ X is the number of 4’s

The simulation is rewritten here for 3 dice to find the probability of X = {0,1,2,3} fours.

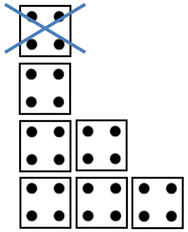
```
//
// Computer simulation of the statistics of throwing three die.
//
#include <iostream>
#include <cstdlib>
int main() {

  const int n = 1000000000;     // Number of trials
  int x[4] = {0};               // Number of "4"s 0,1,2 or 3
  for (int i=0; i<n; i++) {     // Perform n trials
    int m=0;
    for (int j=0; j<3; j++) {   // Loop over the three dice
      int a = rand() % 6 + 1;   // Random outcomes 1,2,3,4,5,6
      if (a==4) m++;            // Count the number of "4"s.
    }
    x[m]++;
  }

  for (int i=0; i<4; i++)
    std::cout << “x=" << i << " P(x)="
    << double(x[i])/n << std::endl;
}
```

```
x=0  P(x)=0.578697
x=1  P(x)=0.347231
x=2  P(x)=0.069445
x=3  P(x)=0.004627
```

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$



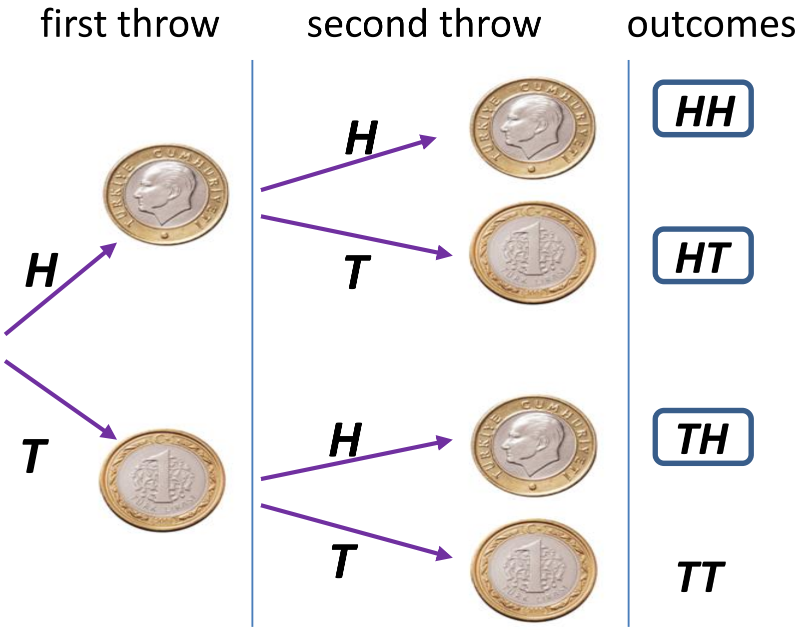
15.00012/216

$$P(A) = \frac{n(A)}{n(S)}$$ *All outcomes are equally likely.*

More on counting sample points:

A coin is thrown twice; what is the probability that *at least one head* occurs?

All possible outcomes can be traced in a tree diagram.



https://en.wikipedia.org/wiki/Tree_diagram_(probability_theory)

Event *A* is the outcome "*at least one head*" which is the collection {*HH*, *HT*, *TH*}

The collection of all possible outcomes is *S* = {*HH*, *HT*, *TH*, *TT*}

Each outcome is equally likely $\Rightarrow$

$$P(A) = \frac{n(A)}{n(S)} = 3/4 = 0.75$$

Counting tool

*Get into the habit of drawing **tree diagrams** to help you trace the paths of different outcomes.*

See also https://en.wikipedia.org/wiki/Event_tree_analysis

The multiplication rule:

If an operation can be performed in n_1 ways, and if for each of these ways a second operation can be performed in n_2 ways, then the two operations can be performed together in $n_1 \times n_2$ ways.

Example: How many possible outcomes are there when a die and a coin are thrown?

Solution: The die can land face-up in n_1= 6 ways; for each of these 6 ways, the coin can land in n_1= 2 ways. Therefore, the total number of possible outcomes is $n_1\, n_2$= 6×2 = **12**.

Draw the tree diagram
and list the outcomes

S = { 1H, 2H, 3H, 4H, 5H, 6H,
1T, 2T, 3T, 4T, 5T, 6T }

$$P(A) = \frac{n(A)}{n(S)}$$

All outcomes are equally likely.

If each "die dot" is assigned a score of 1, a "Tail" is assigned a score of 0 and a "Head" is assigned a score of 1, and scores are summed for the die and coin, what is the probability of scoring more than 3 when the die and coin are thrown?

Solution:

Let *S* represent the set of all possible outcomes

$$S = \{ 1H, 2H, 3H, 4H, 5H, 6H, \\ 1T, 2T, 3T, 4T, 5T, 6T \}$$

The associated scores are:

$$\{ 2, 3, \boxed{4, 5, 6, 7,} \\ 1, 2, 3, \boxed{4, 5, 6} \}$$

Let event *A* represent a score > 3

$$P(A) = \frac{n(A)}{n(S)} = 7 / 12 \approx 0.583$$

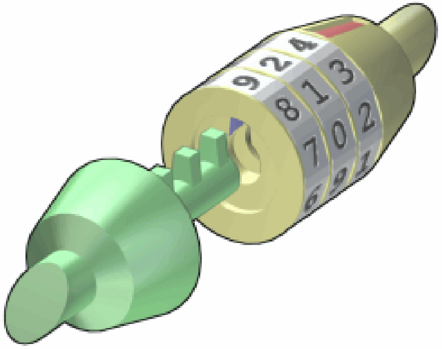
Assuming all outcomes are equally likely.

The generalized multiplication rule:

If an operation can be performed in n_1 *ways, and if for each of these a second operation can be performed in* n_2 *ways, and for each of the first two a third operation can be performed in* n_3 *ways, and so forth, then the sequence of* k *operations can be performed in* $n_1 n_2 n_3 \ldots n_k$ *ways.*

Example

A lock uses the digits 0,1,2,3,4,5,6,7,8,9 on three barrels.
How many possible codes are there?



Combination lock

Bad name! its a *permutation (with repetition) lock*.

Solution

For each of the 10 choices on the first barrel, there are another 10 choices on the second barrel, followed by another 10 choices on the third barrel.

That is 10x10x10 = 1000.

All sequences 000, 001, 002, ... 998, 999.

Theorem: For a set of n elements and selecting k elements; if the **elements can be repeated** then there are n^k permutations.

n=10 digits, k = 3 digits $\Rightarrow 10^3$ = 1000 permutations.

The multiplication rule, applied with some specific rules, gives rise to expressions for permutations and combinations.

Permutations: arrangements where order is counted		**Combinations**: arrangements where order is not counted	
repetition is counted	repetition is not counted	repetition is not counted	repetition is counted
For a set of n elements, select k elements, there are n^k permutations.	For a set of n elements there are $n!$ permutations. Select k elements, there are ${}_nP_k = \frac{n!}{(n-k)!}$ permutations.	For a set of n elements there is **one** combination. Select k elements, there are ${}_nC_k = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!}$ combinations.	For a set of n elements, select k elements $\Rightarrow$ there are $\frac{(n+k-1)!}{k!(n-1)!}$ combinations.

These are shown just to remind you (of your high-school math).

The most important result for this course is:

$${}_nC_k = \frac{n!}{k!(n-k)!} = \binom{n}{k} = \frac{n(n-1)(n-2)....}{k(k-1)(k-2)...1} \quad \text{k terms}$$

Example

Consider that you have bought 6 batteries and charged only 4 of them, but you have forgotten which are charged!

If, in a hurry, you now randomly select 4 batteries to use in a torch, what is the probability that you have selected at least one uncharged battery?

Solution

The number of selected uncharged batteries can be *zero*, *one* or *two*; this is the set of all possible outcomes. Each case has a probability that can be calculated by considering the number of ways to obtain the specific outcome as a ratio of the total number of ways to select the four batteries; that is we can use the definition:

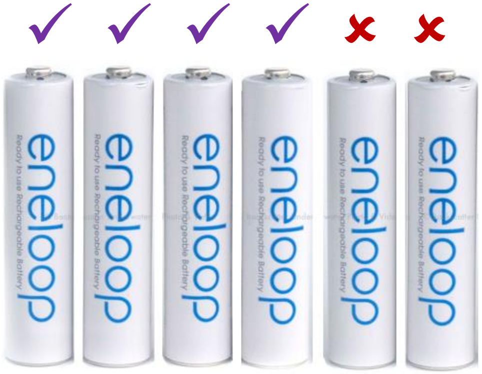
$$P(A) = \frac{n(A)}{n(S)}$$

Since the order of the selected batteries is not important then the "number of ways" is specifically the "number of combinations":

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

$$P(A) = \frac{n(A)}{n(S)} \qquad {}_nC_k = \frac{n!}{k!(n-k)!}$$

The selection of *k* = 0, 1 or 2 uncharged batteries is a two-step process: **select *k* *uncharged* batteries followed by selecting 4-*k* *charged* batteries**; each step has a number of ways (combinations).



The number of ways to select zero (*k*=0) uncharged batteries is: $_2C_0 \times {}_4C_4 = 1$

The number of ways to select one (*k*=1) uncharged battery is: $_2C_1 \times {}_4C_3 = 8$

The number of ways to select two (*k*=2) uncharged batteries is: $_2C_2 \times {}_4C_2 = 6$

The total number of ways to select any 4 batteries from 6 is: $_6C_4 = 15$

The probability that you have selected an uncharged battery (one or both) is:

$$\frac{{}_2C_1 \times {}_4C_3}{{}_6C_4} + \frac{{}_2C_2 \times {}_4C_2}{{}_6C_4}$$

$$= \frac{8}{15} + \frac{6}{15} = \frac{14}{15} \approx 0.933' \ !$$

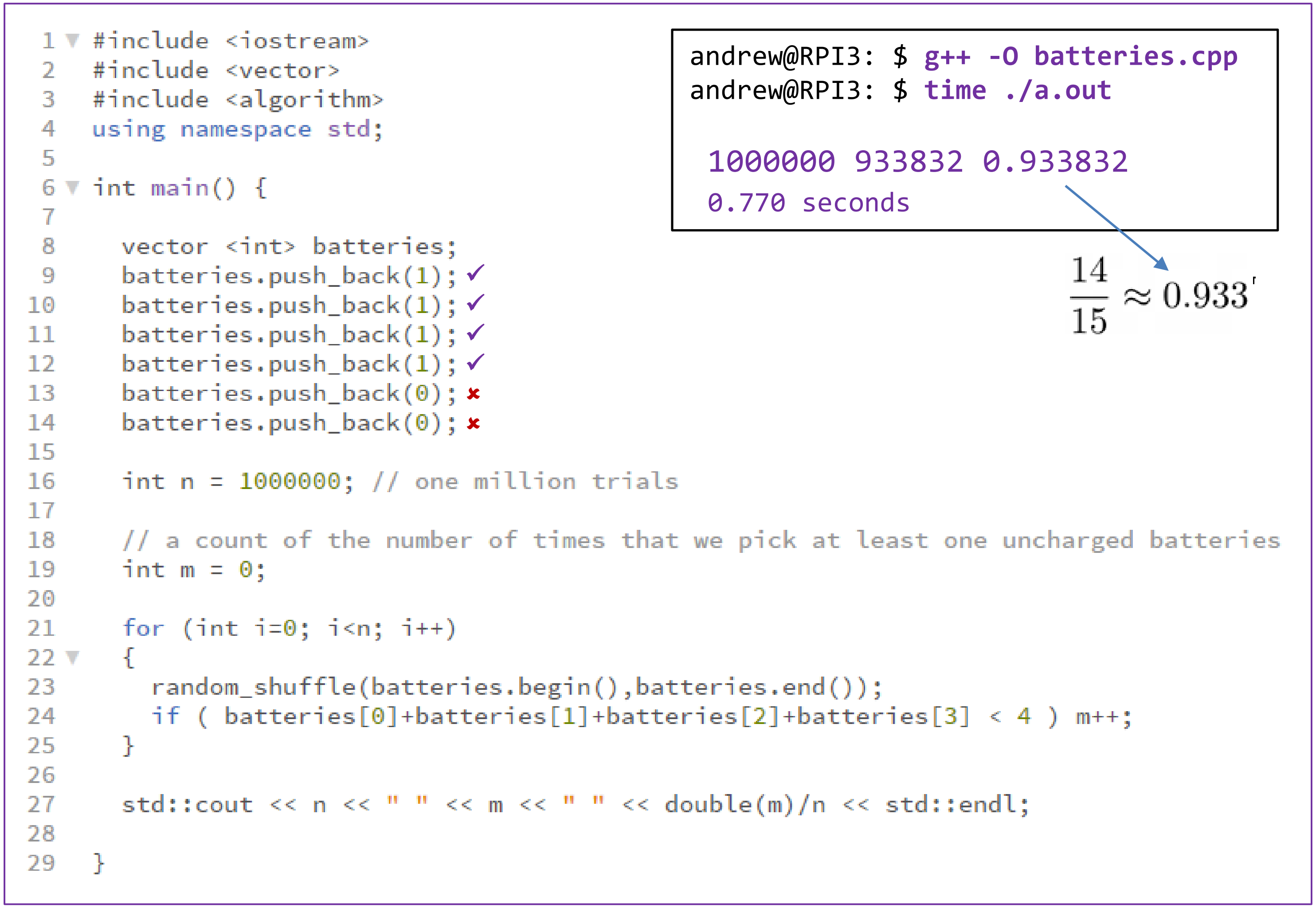
Note that 1 + 8 + 6 = 15 = all the possible outcomes.

This is an example of a class of problems that can be described as "selection without replacement" and follow a hypergeometric distribution. This will be formally studied again later in this course.

COMPUTER PROGRAMMING IS
NOT EXAMINED IN THIS COURSE

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

batteries.cpp



```cpp
#include <iostream>
#include <vector>
#include <algorithm>
using namespace std;

int main() {

  vector <int> batteries;
  batteries.push_back(1); ✓
  batteries.push_back(1); ✓
  batteries.push_back(1); ✓
  batteries.push_back(1); ✓
  batteries.push_back(0); ✗
  batteries.push_back(0); ✗

  int n = 1000000; // one million trials

  // a count of the number of times that we pick at least one uncharged batteries
  int m = 0;

  for (int i=0; i<n; i++)
  {
    random_shuffle(batteries.begin(),batteries.end());
    if ( batteries[0]+batteries[1]+batteries[2]+batteries[3] < 4 ) m++;
  }

  std::cout << n << " " << m << " " << double(m)/n << std::endl;

}
```

```
andrew@RPI3: $ g++ -O batteries.cpp
andrew@RPI3: $ time ./a.out

 1000000 933832 0.933832
 0.770 seconds
```

$$\frac{14}{15} \approx 0.933$$

Summary of today's lecture

- The study of *probability* and *statistics* helps us to make engineering decisions based upon observational data.
  - it's important for society that engineers make good decisions!

- *Randomness* (variability) is the result of complex processes.
  - the world is generally complex.

- Experimentally, the probability of event *A* occuring is $P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$

- Theoretically, we can use *set theory* and basic counting:

$$P(A) = \frac{n(A)}{n(S)}$$ assuming outcomes are equally likely.

- In anyway you define it; we always have $0 \leq P(A) \leq 1$

Homework

General rule: for each hour in class you should study for at least one hour at home (homework).

See **`MAT3026Exercises01.pdf`** for this week's homework exercises.
Homework is not accesses – it is just for your practice.

https://buei.itslearning.com/

- Links to recorded classes
- Links to short "key concepts" videos
- Lecture notes and exercises (not assessed)
- Assignments (assessed!)
- and more ...

Appendix

Appendix

Supplementary notes on *sample spaces*, sets, *events* and *counting*.

Throughout the course we will use *set theory* to describe probabilities with respect to events and sample spaces. If you are not familiar with *sets*, or wish to do some revision, then please study these notes and the sections in the course text book indicated below.

▪ Sample Space	Walpole Section 2.1
▪ Events	Walpole Section 2.2

We will cover these topics again in the lectures as we use them.

2.1 Sample spaces

Definition:
The set of all possible outcomes (the universal set) of a statistical experiment is called the ***sample space*** and is represented by the symbol ***S***. Each outcome in a sample space is called an ***element*** or a ***sample point*** of the sample space.

Example

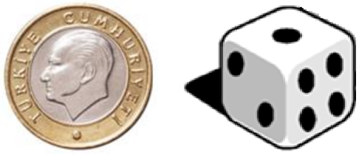
Consider the experiment of tossing a die. If we are interested in the number that appears on the top face, then the sample space is

$$S = \{\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}\}$$

or

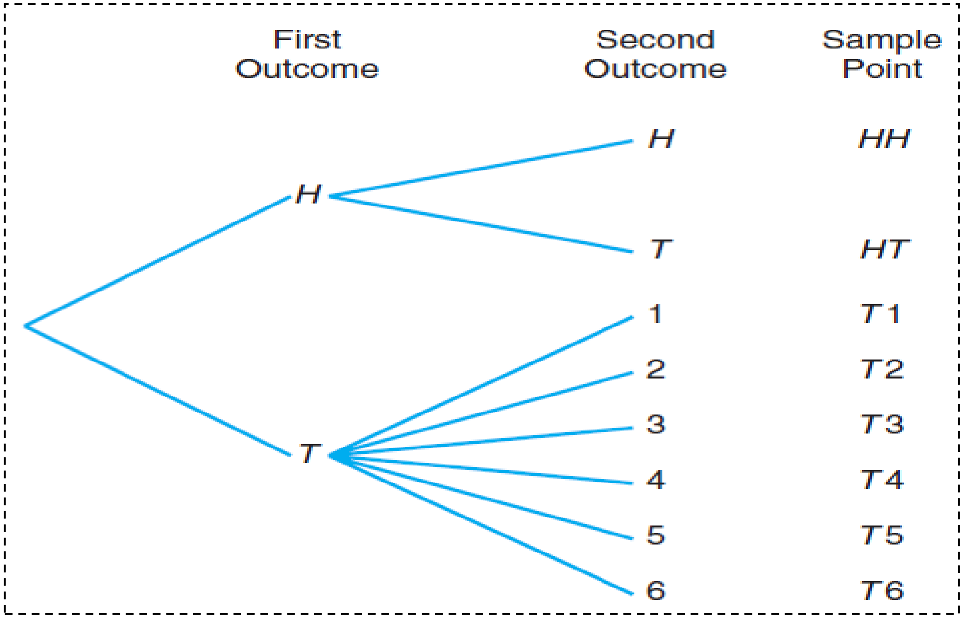
$$S = \{1, 2, 3, 4, 5, 6\}$$

Example : An experiment consists of:



1. toss a coin and then
2. toss it again if a *head* occurs, otherwise throw a die.

All sample points can be found by constructing a *tree diagram*:



Walpole Figure 2.1

By proceeding along all paths, we see that the sample space is:

S = { HH, HT, T1, T2, T3, T4, T5, T6 }

Visualization tool

*A **tree diagram** helps to visualize outcomes of random processes.*

Using a *rule* to define a sample space

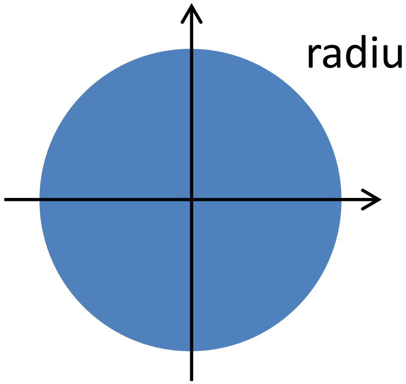
Sample spaces with a large or infinite number of sample points are best described by a *statement* or *rule method*.

Example: $S = \{ x \mid x \text{ is a city with a population over 1 million} \}$

which reads "*S is the set of all x such that x is a city with a population over 1 million.*" The vertical bar means "*such that*."

Example: $S = \{ (x, y) \mid x^2 + y^2 \leq 4 \}$

S is the set of all points (x, y) on the boundary or the interior of a circle of radius 2 with center at the origin



2.2 Events

Definition: An ***event*** is a subset of a sample space.

An event describes a set of outcomes of interest to which we can assign a probability.

Example: consider the sample space of tossing a die:
All possible outcomes are: $S = \{1, 2, 3, 4, 5, 6\}$.

Any subset of S is represented by capital letters such as A, B, C …

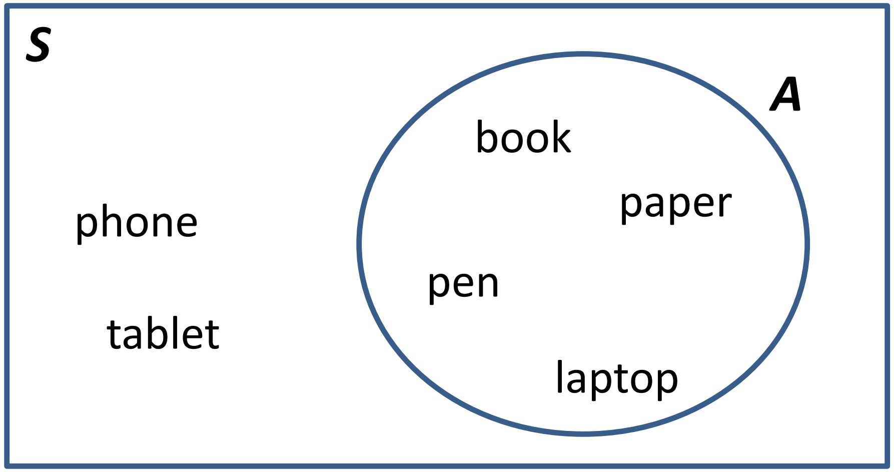
For example $A = \{1, 2, 3, 4\}$ is the event where the outcome is a 1 or a 2 or a 3 or a 4. The probability of this is $P(A) = 4/6$.

Definition: The ***complement*** of an event ***A*** with respect to ***S*** is the subset of all elements of ***S*** that are not in ***A***. This is represented by the symbol ***A'***.

Example

Consider the sample space: ***S*** = {*book*, *phone*, *tablet*, *paper*, *pen*, *laptop*}.

Let set ***A*** = {*book*, *pen*, *laptop*, *paper*}; then set ***A'*** = {*phone*, *tablet*}.



Venn diagram

A' = ***S*** – ***A*** = {phone, tablet}

Visualization tool

*Get into the habit of making **Venn diagrams** to visualize the sets that you are working with.*

Definition: The ***intersection*** of two events A and B, denoted by the symbol $A \cap B$, is the event containing all elements that are common to A and B.

Definition: Two events A and B are ***mutually exclusive*** (or ***disjoint***), if $A \cap B = \emptyset$, that is, if A and B have no elements in common.

Definition: The ***union*** of the two events A and B, denoted by the symbol $A \cup B$, is the event containing all the elements that belong to A or B or both.

Example

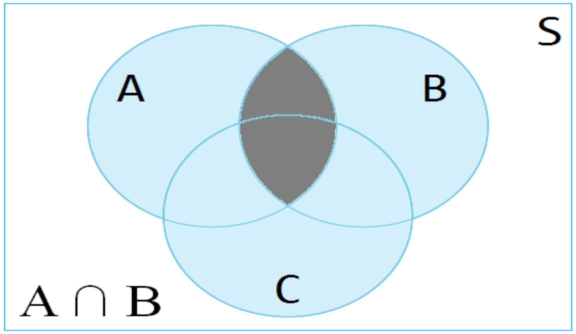
Let $A = \{a, b, c\}$, $B = \{b, c, d, e\}$ and $C = \{d, e, f\}$; then

$A \cap B = \{b, c\}$; $A \cup B = \{a, b, c, d, e\}$; $A \cap C = \emptyset$; $A \cup B \cup C = \{a, b, c, d, e, f\}$.

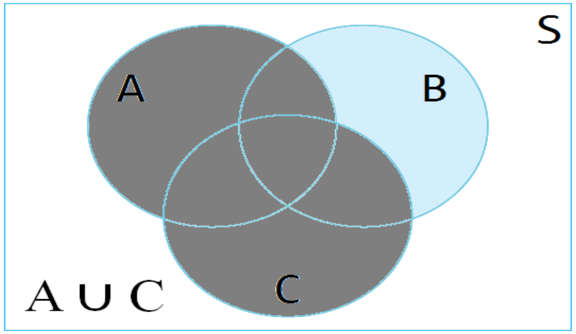
Note: $A \cup B = \{a, b, c\} \cup \{b, c, d, e\} = \{a, \mathbf{b}, \mathbf{c}, \mathbf{b}, \mathbf{c}, d, e\} = \{a, b, c, d, e\}$
That is, the repeated elements are discarded since they do not exist as repeated elements in the sample space.

The relationship between events and the sample space can be illustrated graphically by means of ***Venn diagrams***.

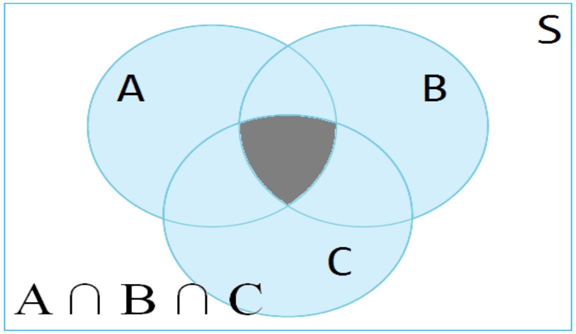
In a Venn diagram we let the sample space be a rectangle and represent events by circles drawn inside the rectangle.



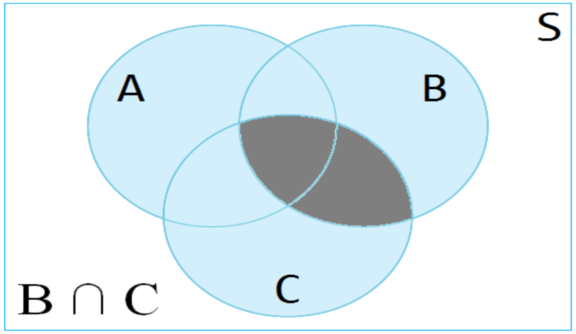
$A \cap B$



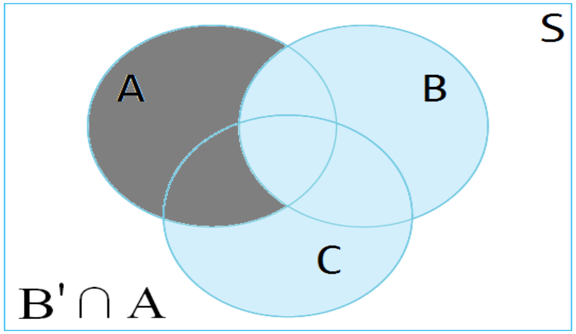
$A \cup C$



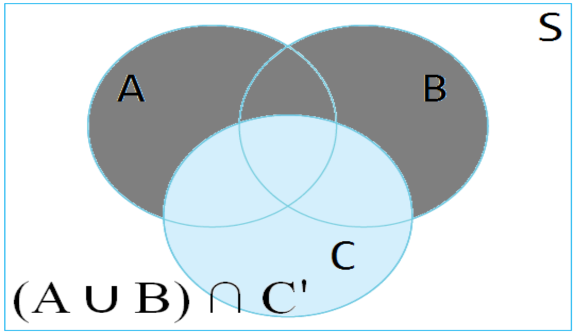
$A \cap B \cap C$



$B \cap C$



$B' \cap A$



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

The following results may easily be verified by means of Venn diagrams (see this week’s exercises).

$$A \cap \phi = \phi$$

$$A \cup \phi = A$$

$$A \cap A' = \phi$$

$$A \cup A' = S$$

$$S' = \phi$$

$$\phi' = S$$

$$(A')' = A$$

$$\left.\begin{aligned}(A \cap B)' &= A' \cup B' \\ (A \cup B)' &= A' \cap B'\end{aligned}\right\} \text{De Morgan's Laws}$$

De Morgan's Laws

The compliment of a union (or an intersection) of two events *A* and *B* is equal to the intersection (or union) of *A’* and *B’*.