Statistics for Life Sciences

Introduction

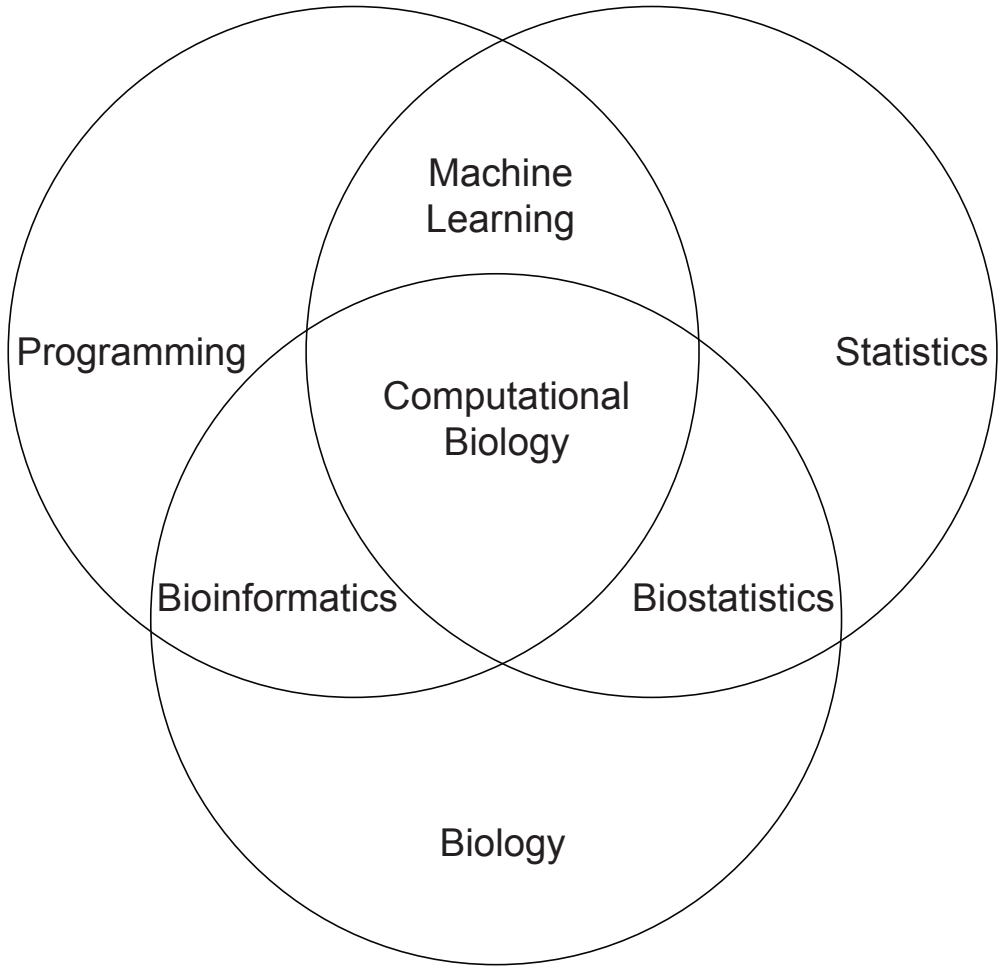
Nathan Harmston

January 10, 2022

More and more data is being generated in the life sciences .... the challenge lies not with the generation of the data, but with the proper experimental design, analysis and interpretation.

- concepts, intuition (sometimes probability is not intuitive), practicality
- Humans are overconfident, arrogant and biased
- we need to avoid making conclusions from limited data, scientists need to use statistics
- implications and limitations of various methods
  - no method is perfect
  - assumptions matter

Statistics for Life Sciences



Statistician vs Biologist

- one aim of this course is to prevent you having conversations like this in the future

`https://www.youtube.com/watch?v=PbODigCZqL8`

- scientists are obsessed with running statistical tests on everything
- *what test did you use? - I don't remember but it gives me a significant p-value*
- one goal of this course is that you never say this to anyone

Syllabus

Week 1	**Data and Probability**
	Summary statistics
	The laws of probability
Week 2	**Data and Probability**
	Visualising data
	Probability distributions
Week 3	**Hypothesis testing – categorical data**
	What is a p-value? – what is a hypothesis test
	Binomial test / Fishers exact test / Chi-squared test
	Testing for the difference of two proportions (theory and simulation)
Week 4	**Hypothesis testing – continuous data**
	t-test
	Wilcoxon signed-rank
	Testing for difference of two means (theory and simulation)
	Confidence intervals

Syllabus

Week 5	**Most common statistical techniques are special cases of linear models**
	Introduction to linear modelling
Week 6	**Most common statistical techniques are special cases of linear models**
	ANOVA
	Likelihood tests
Week 7	**Most common statistical techniques are special cases of linear models**
	Multiple regression
	Logistic regression
Week 8	**Survival analysis**
	Other useful tests / techniques
Week 9	**Experimental design**
Week 10	**Multiple hypothesis testing and power calculations**
Week 11	**Bayesian Statistics** - *in a week*
Week 12	**Lies, damned lies and statistics**
Week 13	**Probabilistic modelling for Life Sciences**

* things are likely to change, be moved around

Assessment

	N	%	Σ
Problem sets	6	5	30
Project	1	30	30
Exam	1	30	30
Participation		10	10
			100

Problem sets

- Most of the questions will involve programming in R
- some will require solving on paper / code
- Coursework to be submitted as a RMarkdown / R script and the associated output files (html)
- Collaboration on problem sets
  - Allowed, but work independently on each problem **before** discussing it
  - Write solutions on your own
  - Acknowledge sources and collaborators

- I prefer to keep deadlines *hard* - but if you need it please come and see me and ask for an extension and it will probably be granted

Project

- process, analyse and interpret some datasets to ask biological questions
- two components of the project related to two distinct datasets

Exam

- will be a combination of questions like what you see for the problem sets, mixed with some more substantial questions
- format / length / etc TBD

Course textbook

We will using **Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking** as a basis for a number of exercises and discussions during the course

probability and statistics (IMHO) definitely fall into the category of just reading about it is completely pointless

its a good introduction / explains a lot of concepts .... but misses out on a few things

Office hours

- Monday afternoons 2-4 and .... another session over zoom (?) will be first come, first served! I may end up working with two or more of you at the same time!
- These are there to help you with material from the whole course

Ask questions

You're an muppet if you have a question and don't ask it, so don't be a muppet

ASK THE QUESTION

If you have a question on something - you're not a muppet and you're probably not the only one - don't be selfish - think of the class

ASK THE QUESTION

If I talk about something you don't think I've described well enough (or I haven't described it at all) - stop me from being a muppet*

ASK THE QUESTION

* this is the first time this course has been ran - there will be parts of this course when I will be a muppet

Questions?

Statistics and probability are not intuitive

statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write

H.G. Wells

What can we do with statistics?

- describe - simplify complexity
- decide - make decisions based on uncertain data
- predict - predict a new situation based on previous one

Statistics and probability are not intuitive

- we tend to jump to conclusions - we over generalise
- we tend to be overconfident

Motulsky CHAPTER ONE

Statistics and probability are not intuitive

Answer these questions and pick a range which you think has a 90% chance of containing the correct answer?

- Martin Luther King Jrs age at death?
- length of the Nile in kilometers?
- number of countries in OPEC?
- Number of books in the new testament?
- Diameter of the moon in miles?
- Weight of a Boeing 747 in kilograms?
- Year Mozart was born?
- Gestation period of an Asian elephant in days?
- Distance from London to Tokyo in kilometers?
- Deepest known point in the ocean in kilometers?
- Year I was born?

Statistics and probability are not intuitive

Answer these questions and pick a range which you think has a 90% chance of containing the correct answer?

- Martin Luther King Jrs age at death?
- length of the Nile in kilometers?
- number of countries in OPEC?
- Number of books in the new testament?
- Diameter of the moon in miles?
- Weight of a Boeing 747 in kilograms?
- Year Mozart was born?
- Gestation period of an Asian elephant in days?
- Distance from London to Tokyo in kilometers?
- Deepest known point in the ocean in kilometers?
- Year I was born?

99% of people tested were overconfident - Russo and Schoemaker 1989

Statistics and probability are not intuitive

- we tend to jump to conclusions - we over generalise (**hypothesis testing** & **model comparison**)
- we tend to be overconfident (**confidence intervals)**
- we see patterns in random data (hot hands fallacy, if we go looking for something we're biased to see it)
- we don't realise that coincidences are common
- we don't expect variability to depend on sample size (**power calculations**)
  - random variation can have a bigger effect on averages within small groups than within large groups
- we find it hard to combine probabilities (**probability** & **Bayesian statistics**)
- we don't do Bayesian calculations intuitively (**probability** & **Bayesian statistics**)
- we are fooled by multiple comparisons (**hypothesis testing** & **multiple hypothesis testing**)
- we tend to ignore alternative explanations (**experimental design**
- we are fooled by **regression to the mean**
- we let our biases determine how we interpret data

which is greater? the number of six-letter English words having *n* as their fifth letter or the number of six-letter English words ending in *ing*?

which is greater? the number of six-letter English words having *n* as their fifth letter or the number of six-letter English words ending in *ing*?

availability bias

Randomness

really we want to **describe** and **understand** randomness

we want to ask whether what we observe / end up with is different to what we'd expect by chance (observed vs. expected)

- average
- variability
- significance
- quantify chance

PROBABILITY

- study of randomness
- provides the foundation for statistics
- tools to handle uncertainty

Probability

probability was really worked on because of **gambling** and for **insurance**

a lot of examples always involve coins, dice, cards etc. (and sometimes roulette), and urns (so many urns)

Problem of the points - Pascal and Fermat - 1600s

John Law

Edmund Halley

Probability

experiment - something happens - flip a coin, draw 5 cards from a deck, roll some dice, pull some balls out of an urn

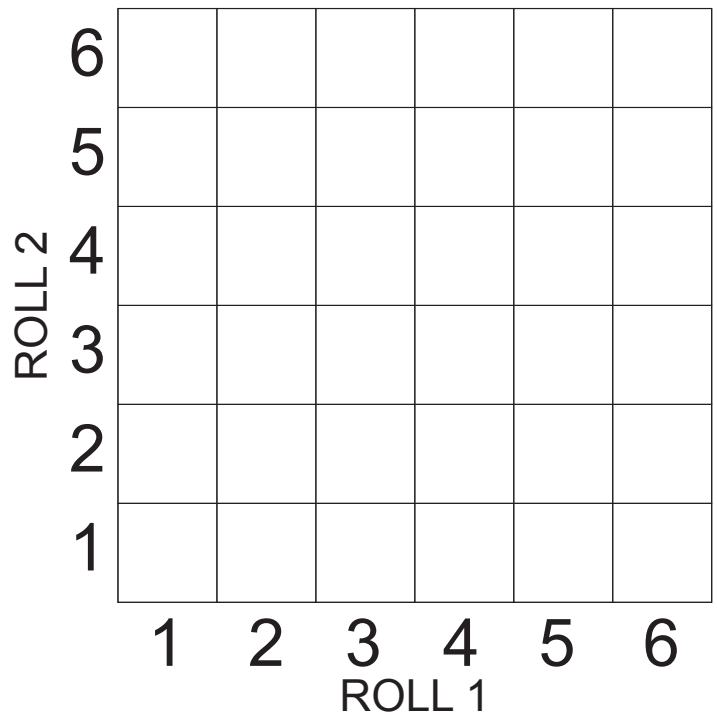
sample space - Ω - set of outcomes

- mutually exclusive
- exhaustive

Roll a dice twice

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

Sample space



Probability

each subset of the sample space -> **events** - each event is associated with a probability

P(A) - probability that event A will occur
P(B) - probability that event B will occur

Two interpretations of probability

Subjective probability - Bayesian

degree of belief in a particular proposition/statement

inside your head

Empirical frequency - frequentist

do the experiment multiple times and count how often the event occurs?

calculate probability from the relative frequency of events

law of large numbers

Motulsky CHAPTER TWO

Law of large numbers

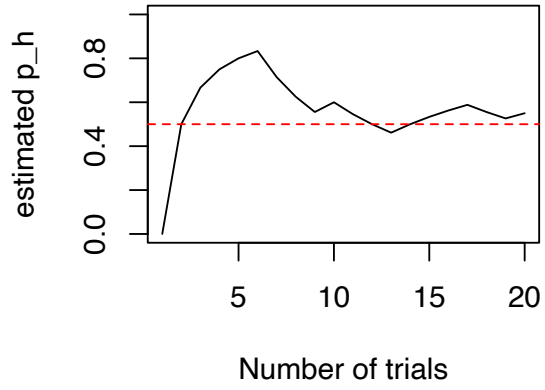
empirical probability will approach the true probability as the sample size increases

Simulate ....

simulate a large number of coin tosses, looking at our estimate of the probability of heads after each flip;
P(H) = 0.5

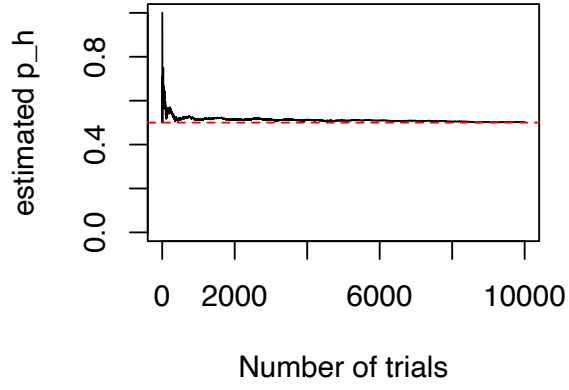
Law of large numbers

empirical probability will approach the true probability as the sample size increases

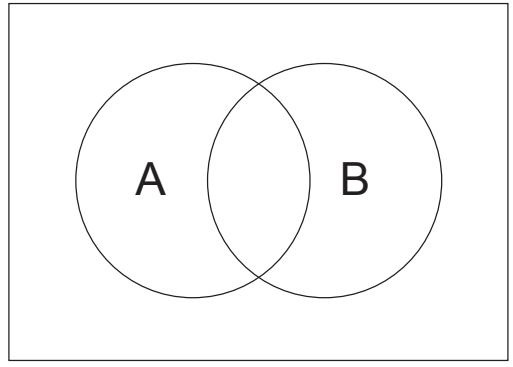


Simulate ....

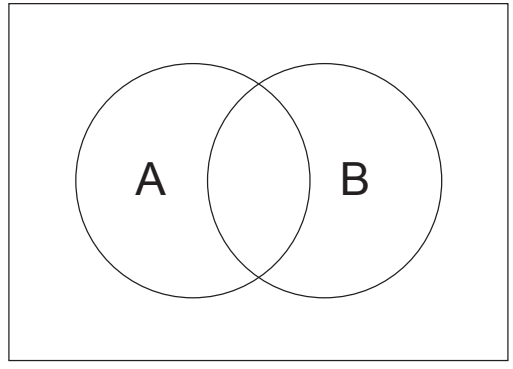
simulate a large number of coin tosses, looking at our estimate of the probability of heads after each flip;
P(H) = 0.5



Venn diagrams - playing with sets



intersection - **and** - $\cap$

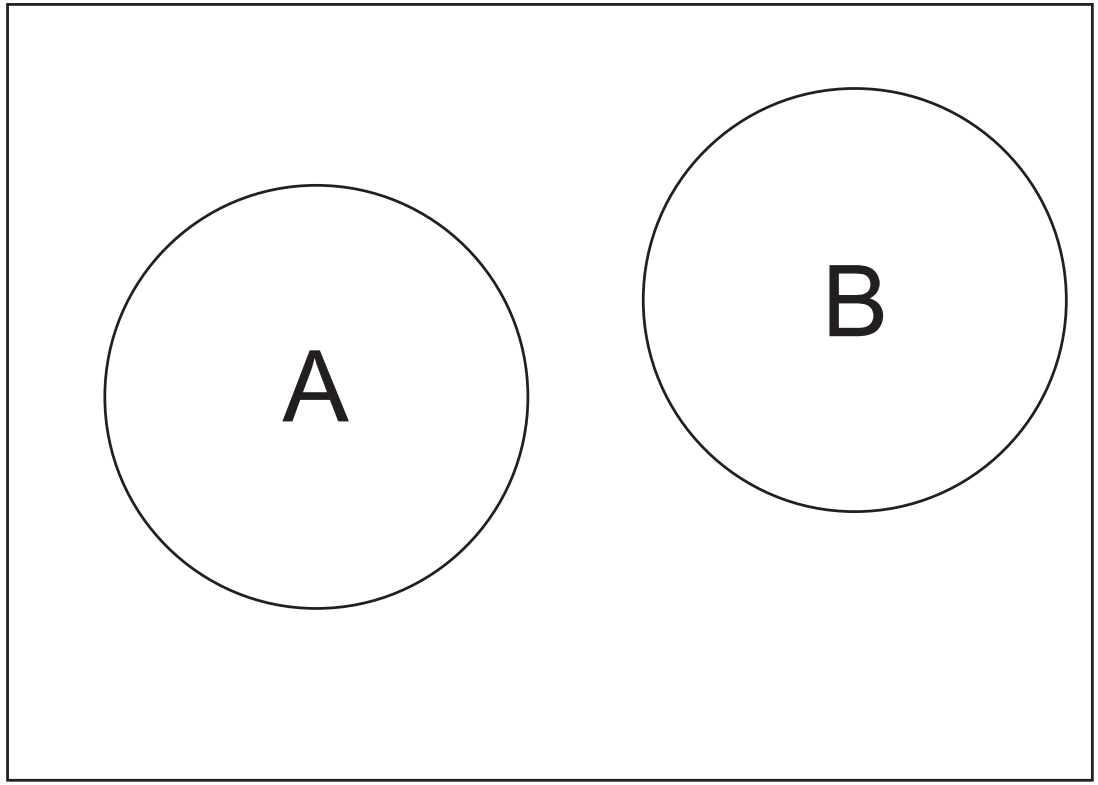


union - **or** - $\cup$

Laws of probability

- non-negativity = $P(A) \geq 0$
- normalisation = $P(\Omega) = 1$
- additivity = if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

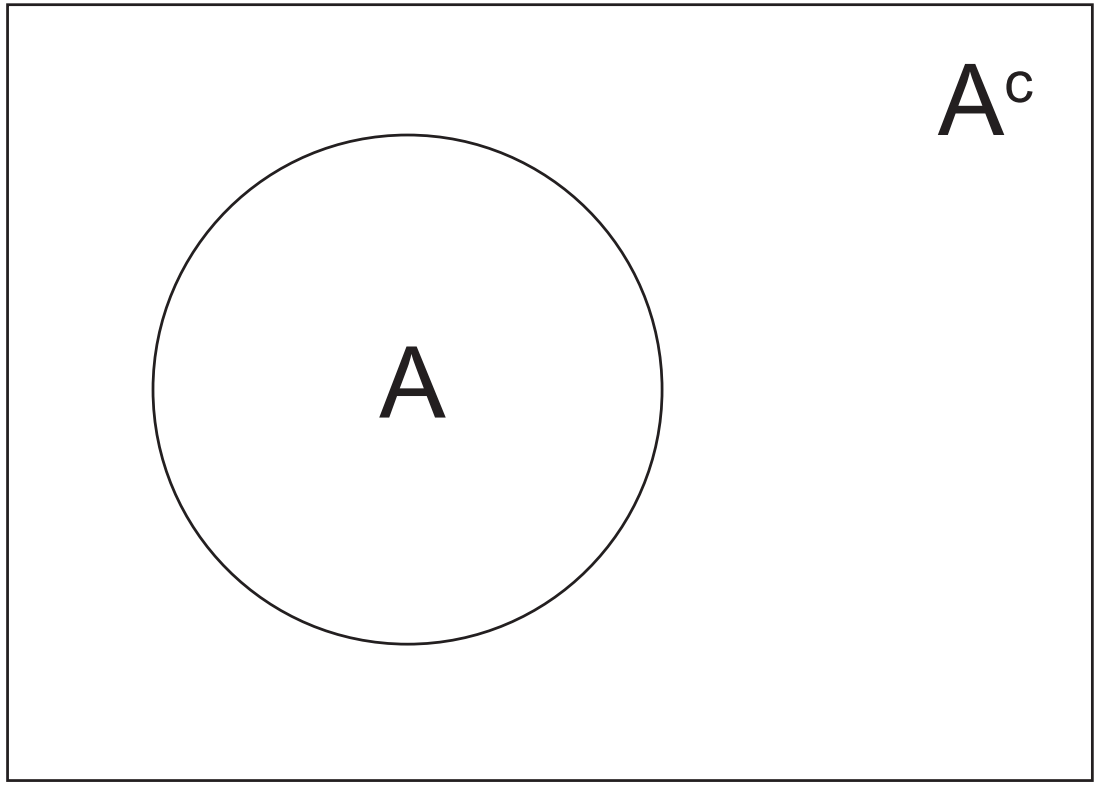
Additivity



how do you show that $P(A) \leq 1$?

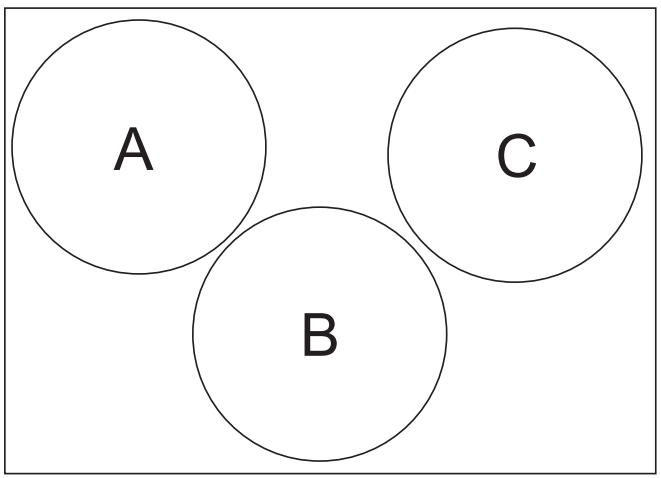
The complement rule

$$P(A^C) = 1 - P(A)$$



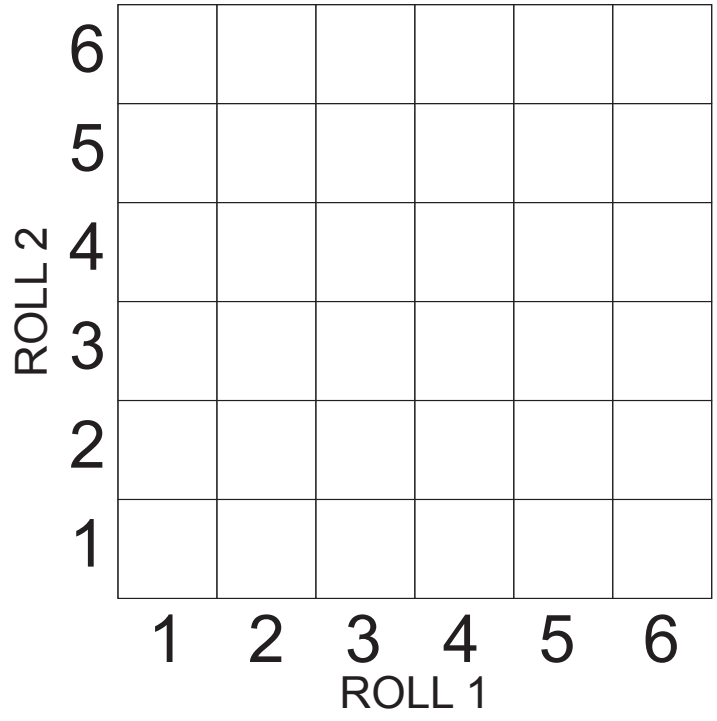
Additivity for more than 2 sets?

$P(A \cup B \cup C)$



Lots of the time probability is just simply just counting

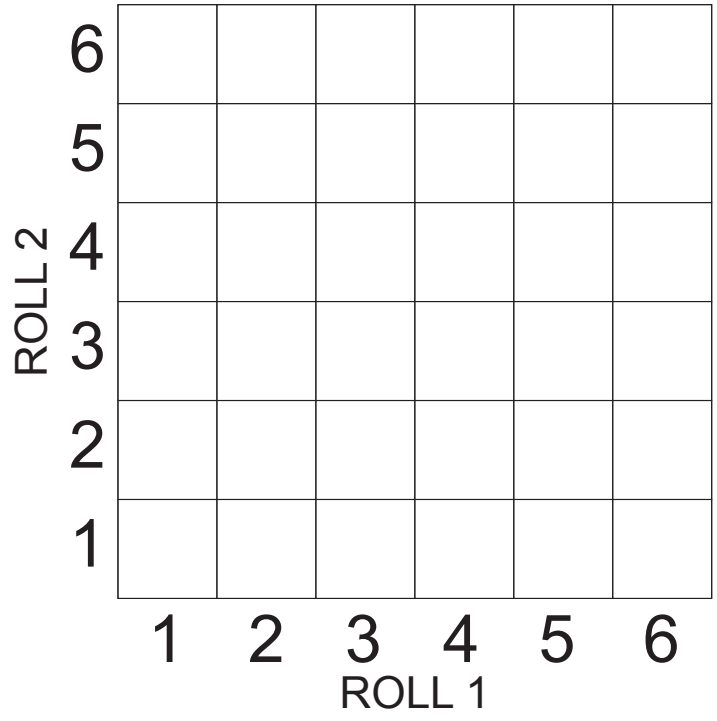
Sample space



$$P((X, Y) = (1,1) \text{ or } (1,2))$$

Lots of the time probability is just simply just counting

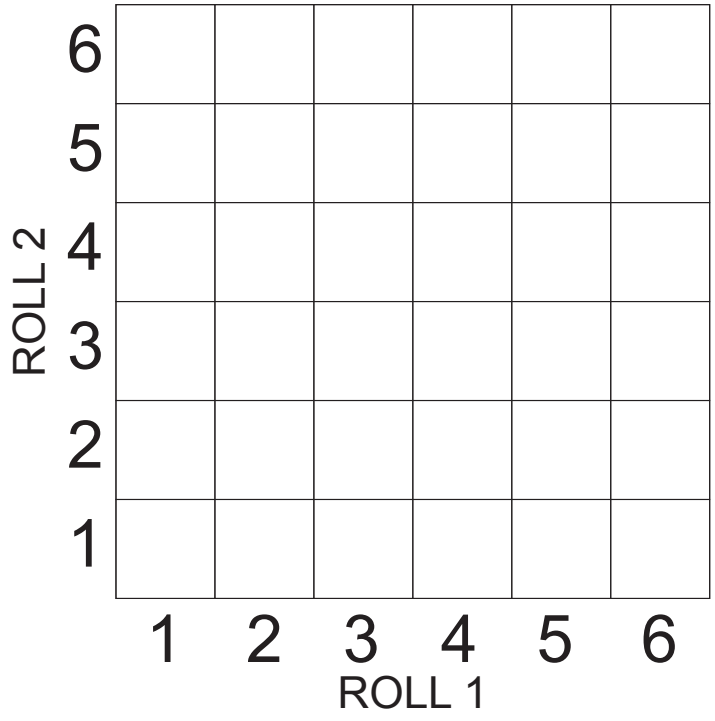
Sample space



$P(X=1)$

Lots of the time probability is just simply just counting

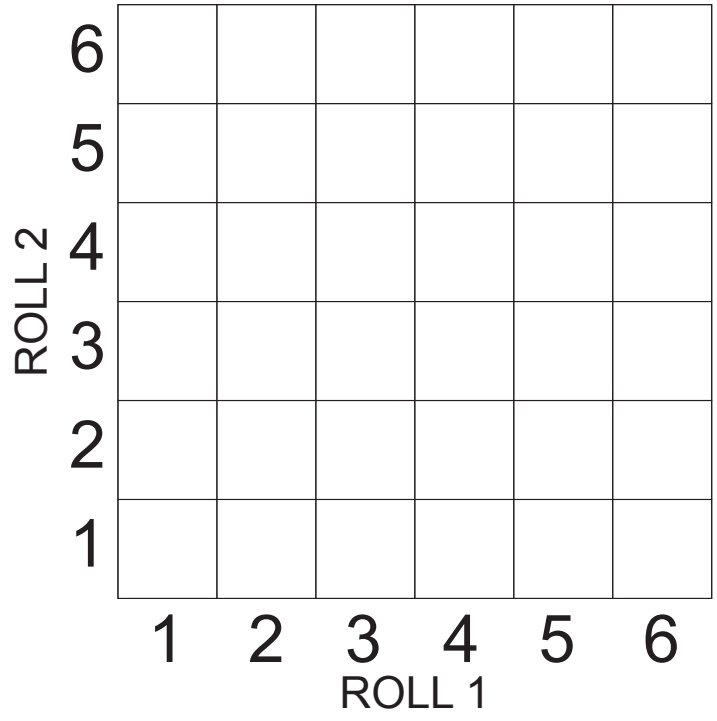
Sample space



P(X+Y is even)

Lots of the time probability is just simply just counting

Sample space



$$P(\min(X,Y) = 3)$$

Making decisions using probability

make decisions - reason in the face of uncertainty

Roll a dice ...

1. you receive 1 SGD if the number of dots ≤ 3
2. you receive 2 SGD if the number of dots ≤ 2

which option would you choose and why?

Making decisions using probability

Roll two dice ...

- pick a number between 2 and 12 - if you guess correctly you win 100 SGD

which number would you pick and why?

Making decisions using probability - Jelly beans

nine white and one red jelly beans

93 white beans and seven jelly beans

You win 100 SGD if you pick a red bean

which one do you pick?

and yet ....

2/3 of people prefer to choose from the larger bowl

the large one has more chances to win

lots of people make the *illogical choice* - human beings are muppets

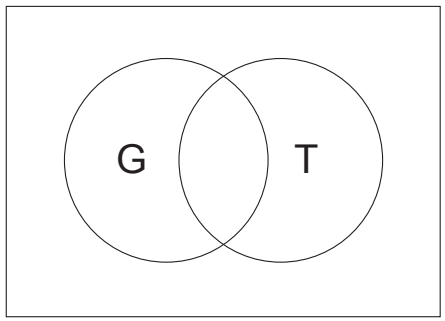
Denes, Raj and Epstein 1994

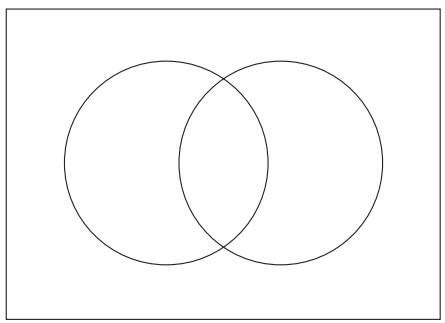
Do geniuses drink tea?

50% of people are geniuses
80% of people drink tea
40% of people are geniuses and drink tea

what is the probability that someone is not a genius **and** is not a tea-drinker?

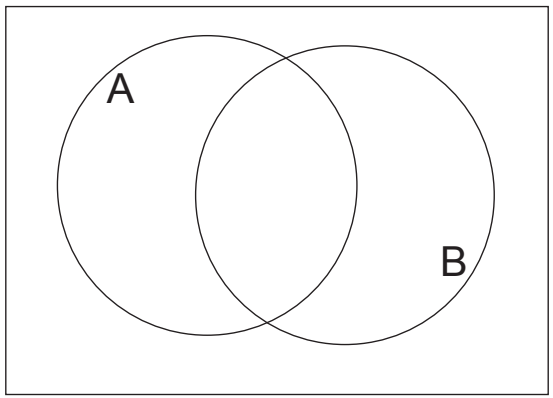
Do geniuses drink tea?





Conditional probability

$P(A|B)$ - probability that event A will occur given that event B has occurred



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B) > 0$$

Conditional probability

$P(A|B)$ - probability that event A will occur given that event B has occurred

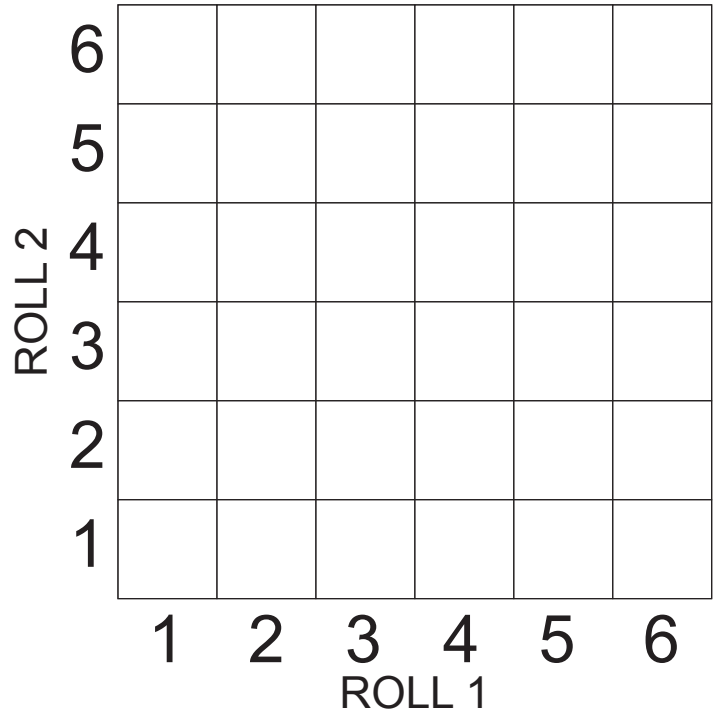
$$P(A \cap B) = P(B)P(A|B)$$

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

$$P(B) > 0$$

Sometimes probability is just simply just counting

Sample space



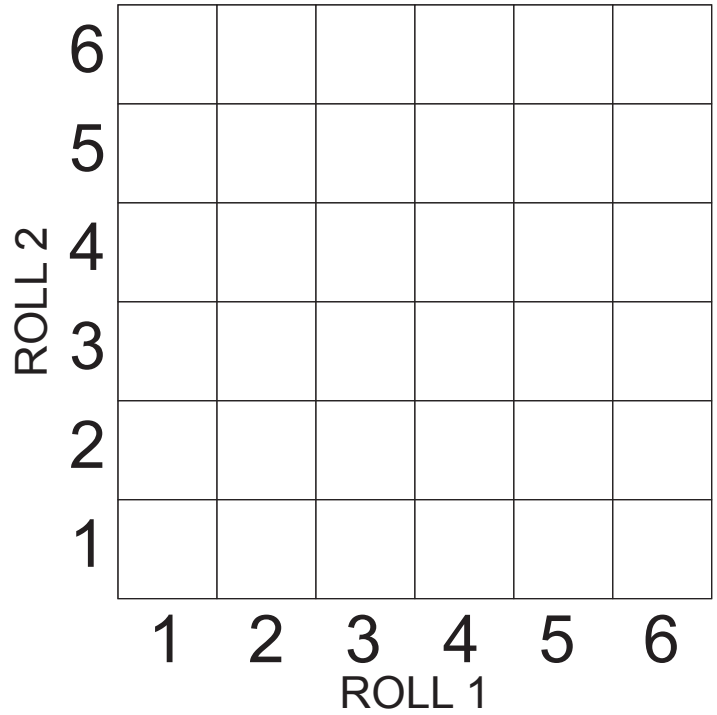
$$B = \min(X,Y) = 2$$

$$M = \max(X,Y)$$

$$P(M = 1 \mid B)$$

Sometimes probability is just simply just counting

Sample space



$$B = \min(X,Y) = 2$$

$$M = \max(X,Y)$$

$$P(M = 2 \mid B)$$

Bayes' theorem

Inverse problems

given P(A|B) what is P(B|A)?

reverse conditional probabilities

Baye's rule

$$P(A|B) = \frac{P(B|A)}{P(A)} P(B)$$

We don't do Bayesian calculations intuitively

Question

Only a tiny fraction (0.1%) of the people have a disease D. A test for this disease is highly accurate but not quite perfect. It correctly identifies 99% of patients with the disease but also incorrectly concludes that 1% of the noninfected samples have the disease. When this test identifies a blood sample as having HIV present, **if you have a positive result what is the chance that you have the disease**?

Bayes' theorem

- P(D) - probability of having the disease -
- P(T) - probability of having a positive test
  - $P(T) = \Sigma P(a_i) x P(b|a_i)$ - law of total probability
- P(D | T) = what we want to find - probability of disease given a positive test
- P(T | D) = probability of a positive test given you have the disease

Bayes' theorem

$$P(D|T) = 0.09$$

the interpretation of the result depends on the fraction of the population that has the disease

Questions?

- Friday - probability, independence, random variables, Binomial distribution
- Next week - handling and visualising data and probability distributions