

Probability

Michael Noonan

Biol 520C: Statistical modelling for biological data

1. Review
2. Probability Theory 101
3. Bayes' Theorem
4. Probability Distributions

Review

Last lecture we covered how simple linear regression models fit by least squares provides a formal description of the deterministic components of a system where y is proportional to x :

$$y_i = \beta_0 + x_i\beta_1$$

Minimising the sum of the squares provided a path forward for the deterministic part, but provided no description of our model's stochastic component.

This led us to approaching the problem as probabilists.



How does that help?

Probability Theory 101

If you were to look around UBC's libraries you'd probably find dozens if not hundreds of books on probability theory. There's no way I could teach you all of this material in a single lecture.

I'm not trying to teach you 'probability theory', but I am trying to give you those pieces you'll need to understand the concepts relevant to this course.

Hopefully this will also motivate you to take a deeper dive into probability theory and probability distributions outside of this course.

Biological data are very noisy, and full of 'randomness'.

Being a good modeller means being able to understand not only the deterministic part of a model, but also the stochastic part.

Most introductory statistics courses teach you methods that assume the randomness of a process is normally distributed, in many cases this is totally fine, in many others though this is not an appropriate assumption.

In order to make sense of a system's stochasticity, we need to rely on probability distributions. In order to work with probability distributions, we need to understand some probability theory.

In probability theory we are concerned with the occurrence of random events.

More specifically, we're interested not in single outcomes, but average outcomes over a large number of replicates (flipping coins, rolling die, picking names from a hat, etc...)

What group of people have a lot of experience with the outcomes of random chance events? Gamblers.



Pascal Source: Wikipedia



de Fermat Source: Wikipedia

In the mid 1600s when a professional gambler asked French mathematician Pierre de Fermat why if he bet on rolling at least one six in four throws of a die he won in the long term, whereas betting on throwing at least one double-six in 24 throws of two dice resulted in his losing on average.

de Fermat worked with Blaise Pascal to show mathematically why this was the case...

de Fermat worked out that

$$\begin{aligned}\text{Prob. of one 6 in 4 throws} &= 1 - \text{Prob. of no 6 in 4 throws} \\ &= 1 - (5/6)^4 \\ &= 0.518 \text{ (i.e., winning on average)}\end{aligned}$$

Whereas

$$\begin{aligned}\text{Prob. of 6-6 in 24} &= 1 - \text{Prob. of no 6-6 in 24} \\ &= 1 - (35/36)^{24} \\ &= 0.491 \text{ (i.e., losing on average)}\end{aligned}$$

...and this work became the foundation of modern probability theory.

In probability theory we are concerned with the occurrence of random events.

(Think of an event as the outcome of an experiment)

We write this:

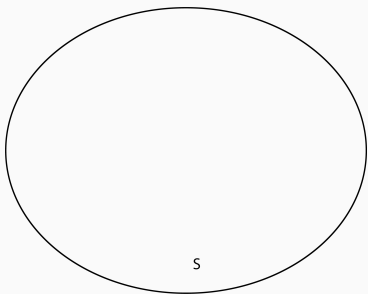
$\Pr\{A\}$ = Probability that event A occurs,

$\Pr\{B\}$ = Probability that event B occurs,

etc...

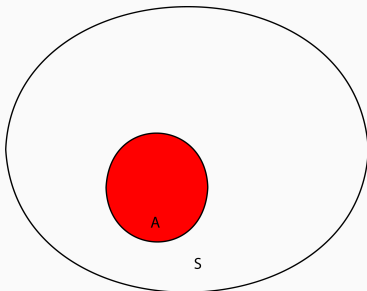
Let's say S is the collection of all possible outcomes of our 'experiment' (sides on a coin, numbers on a die, possible ages, whatever)

This collection of outcomes is termed the 'sample space', the sum of all the probabilities in the sample space is 1 ($\Pr\{S\} = 1$)



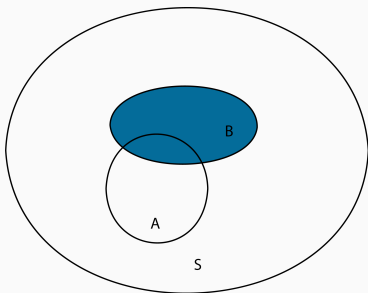
For tossing a single six-sided die, the sample space is $\{1, 2, 3, 4, 5, 6\}$.

We carry out our experiment and we observe event 'A'



$$\begin{aligned}\Pr\{A\} &= \text{Probability of event } A \\ &= (\text{area of } A) / (\text{area of } S)\end{aligned}$$

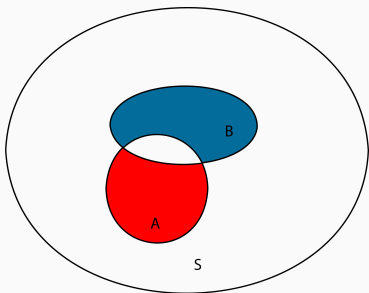
We carry out our experiment again and we observe event 'B'



$$\Pr\{B\} = (\text{area of } B) / (\text{area of } S)$$



What about the probability of either 'A' or 'B' ?



$$\Pr\{A \text{ or } B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \text{ and } B\}$$

Note: more formally the $\Pr\{A \text{ and } B\}$ is denoted as $\Pr\{A, B\}$

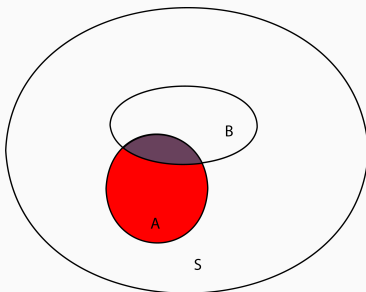
What about the probability of 'B' given 'A' occurred?

This is termed conditional probability (i.e., the probability of an event under the condition that another event occurred)

Events follow each other all the time in reality.

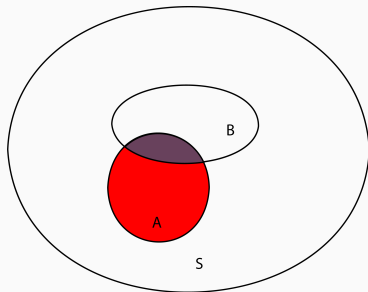
The probability of event $B = \Pr\{B\} = (\text{area of } B) / (\text{area of } S)$, but if we know that 'A' happened...

$\Pr\{B \text{ given that } A \text{ occurred}\} = (\text{area common to } A \text{ and } B) / (\text{area of } A)$

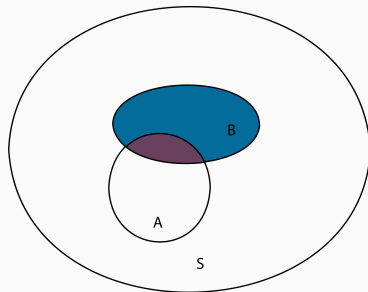


or more formally, $\Pr\{B|A\} = \Pr\{A,B\}/\Pr\{A\}$

$$\Pr\{B|A\} = \Pr\{A,B\}/\Pr\{A\}$$



$$\Pr\{A|B\} = \Pr\{A,B\}/\Pr\{B\}$$



Events are independent when knowing that one occurs does nothing to change our idea about the probability of another event occurring.

If two events are independent of one another, then

$$\Pr\{A|B\} = \Pr\{A\} \quad \text{and} \quad \Pr\{B|A\} = \Pr\{B\}$$

And if we remember that

$$\Pr\{A|B\} = \Pr\{A,B\}/\Pr\{B\} \quad \text{and} \quad \Pr\{B|A\} = \Pr\{A,B\}/\Pr\{A\}$$

then

$$\Pr\{A|B\} = \Pr\{A,B\}/\Pr\{B\}$$

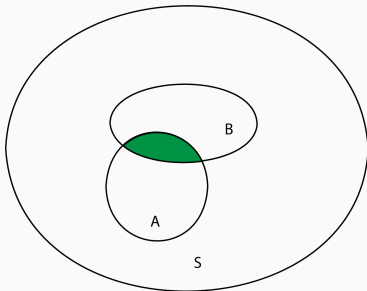
$$\Pr\{A\} = \Pr\{A,B\}/\Pr\{B\}$$

$$\Pr\{A,B\} = \Pr\{A\}\Pr\{B\}$$

$\Pr\{A,B\}$ is referred to as the joint probability of A and B.

You'll often see this written as $\Pr\{A \cap B\}$

The \cap symbol refers to the 'intersection' of A and B.



Bayes' Theorem

Wikipedia: In probability theory and statistics, Bayes' theorem (alternatively Bayes's theorem, Bayes's law or Bayes's rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.



But what does that actually mean?

From earlier, we had:

$$\Pr\{A|B\} = \Pr\{A,B\}/\Pr\{B\}$$

Rearranging this, we get:

$$\Pr\{A,B\} = \Pr\{A|B\} \Pr\{B\}$$

Remembering that the probability of B given A is given by:

$$\Pr\{B|A\} = \Pr\{A,B\}/\Pr\{A\}$$

We can get to:

$$\Pr\{B|A\} = \Pr\{A|B\} \Pr\{B\}/\Pr\{A\}$$

The mathematical description of Bayes' Theorem is given as:

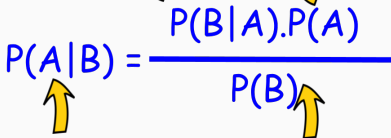
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

LIKELIHOOD

The probability of "B" being True, given "A" is True

PRIOR

The probability "A" being True. This is the knowledge.


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

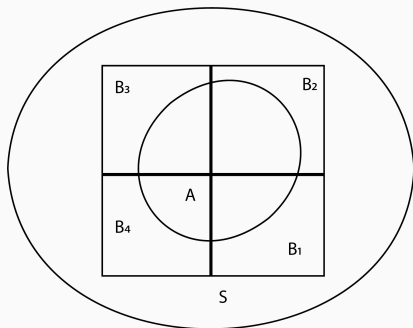
POSTERIOR

The probability of "A" being True, given "B" is True

MARGINALIZATION

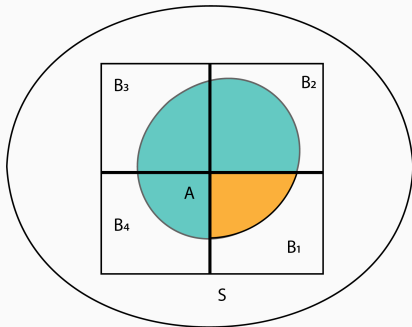
The probability "B" being True.

Bayes' theorem is most useful when there are multiple, exclusive possible outcomes, B_1, B_2, \dots, B_N , and one must occur when A occurs.



$$Pr\{B_i|A\} = \frac{Pr\{A|B_i\}Pr\{B_i\}}{\sum_{j=1}^N Pr\{A|B_j\}Pr\{B_j\}}$$

$$Pr\{B_1|A\} = \frac{Pr\{A|B_1\}Pr\{B_1\}}{\sum_{j=1}^4 Pr\{A|B_j\}Pr\{B_j\}}$$



Question: I've flipped two coins, and I tell you 1 came up heads. What's the probability the other flip was heads?

Approach #1: If each coin flip is independent, and heads/tails are equally probable, then:

$$\Pr\{\text{Heads}_2 \mid \text{Heads}_1\} = \Pr\{\text{Heads}_2\} = 1/2$$

Approach #2: There are 4 possible outcomes: {HH, HT, TH, TT}. If 1 flip is heads, TT is impossible. If each combination is equally likely, then:

$$\Pr\{\text{HH}\} = 1/3$$

Both approaches make intuitive sense, but both can't be right.

The challenge is how to use the information I've given to you that 1 flip is heads.

What we want to know is:

$$\Pr\{HH \mid \text{knowing one flip is H}\} = \frac{\Pr\{HH, \text{ knowing one flip is H}\}}{\Pr\{\text{knowing one flip is H}\}}$$

Allowing all 4 sets of possible outcomes, we have:

Flip Results	Prior probability	Pr{H given flip results}
HH	1/4	1
HT	1/4	1/2
TH	1/4	1/2
TT	1/4	0

Flip Results	Prior probability	Pr{H given flip results}
HH	1/4	1
HT	1/4	1/2
TH	1/4	1/2
TT	1/4	0

Next we need to calculate the joint probability of each outcome and you knowing I flipped 1 heads:

$$\Pr\{\text{HH}, H\} = \Pr\{\text{HH}\} \Pr\{H \text{ given flip result}\} = 1/4 \times 1 = 1/4$$

$$\Pr\{\text{HT}, H\} = 1/4 \times 1/2 = 1/8$$

$$\Pr\{\text{TH}, H\} = 1/4 \times 1/2 = 1/8$$

$$\Pr\{\text{TT}, H\} = 1/4 \times 0 = 0$$

So $\Pr\{\text{of knowing 1 flip is heads}\} = 1/4 + 1/8 + 1/8 = 1/2$

$$\Pr\{HH \mid \text{knowing one flip is H}\} = \frac{\Pr\{HH, \text{ knowing one flip is H}\}}{\Pr\{\text{knowing one flip is H}\}}$$

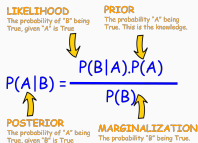
$$\Pr\{HH, \text{ knowing one flip is H}\} = 1/4$$

$$\Pr\{\text{of knowing 1 flip is H}\} = 1/2$$

$$\Pr\{HH \mid \text{knowing one flip is H}\} = \frac{1/4}{1/2} = 1/2$$

So, our first approach from earlier was correct.

You'll often see people argue that the strength of Bayesian methods is the ability to make use of 'prior' information (e.g., previously collected data).



LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. That is the knowledge.

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

POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

$$Pr\{B_i|A\} = \frac{Pr\{A|B_i\}Pr\{B_i\}}{\sum_{j=1}^N Pr\{A|B_j\}Pr\{B_j\}}$$

But in practice, people usually calculate the prior based on some existing dataset, so you could skip this if you just used the original data.

A lot of the time you will also see people using 'flat uninformative prior', which means the prior isn't really doing anything meaningful.

The biggest benefit (in my opinion) comes from being able to use computer algorithms to calculate the denominator (marginal).

Probability Distributions

We're going to finish by briefly reviewing a number of commonly used probability distributions.

This list is not exhaustive, but it should be sufficient for allowing you to calculate $\Pr\{\text{model}|\text{data}\}$ and $\Pr\{\text{data}|\text{model}\}$ for many ecological scenarios.

You **do not** need to memorise the formulae, but you should be able to recognise them, and understand their basic properties and use cases.

The binomial distribution describes the probability of obtaining k yes/no successes in a sample of size n , or in other words, the distribution of the number of successful trials among a defined number of trials.

Parameters: n and p

Type: Discrete

Biological scenarios: Mark recapture data, live vs dead survival data, killed by a predator or not, yes/no behavioural outcomes, anything with a discrete yes/no outcome.

PMF: $\binom{n}{k} p^k (1-p)^{n-k}$

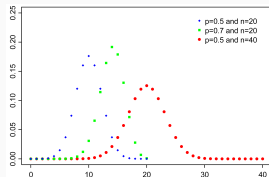
where

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

Range: discrete ($0 \leq x \leq n$)

Mean: np

Variance: $np(1-p)$



The Poisson distribution describes the probability of a given number of events occurring in a fixed interval of time or space.

Parameters: λ

Type: Discrete

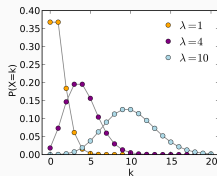
Biological scenarios: Counts of a species per unit time, the number of mutations on a strand of DNA per unit length, number of births/deaths per year in a given age group, prey caught per unit time.

$$\text{PMF: } \Pr(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Range: discrete $(0, \infty)$

Mean: λ

Variance: λ



Source: Wikipedia

The negative binomial distribution describes the number of *failures* in a sequence of independent and identically distributed trials.

Parameters: p Probability per trial,
 k Overdispersion parameter

Type: Discrete

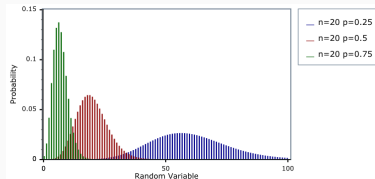
Biological scenarios: Same as the Poisson distribution, but allowing for more heterogeneity because variance \neq mean.

$$\text{PMF: } \frac{\Gamma(k+r)}{k! \cdot \Gamma(r)} p^k (1-p)^r$$

Range: discrete ($x \geq 0$)

$$\text{Mean: } \frac{pr}{1-p}$$

$$\text{Variance: } \frac{pr}{(1-p)^2}$$



The Gaussian (or normal) distribution is a continuous, symmetrical distribution that applies frequently in practice.

Parameters: μ and σ

Type: Continuous

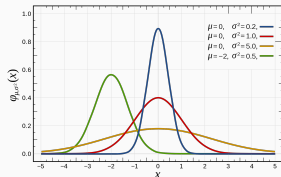
Biological scenarios: Many.
Almost any measurement that is continuous and symmetrically distributed.

$$\text{PDF: } \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Range: $(-\infty, \infty)$

Mean: μ

Variance: σ^2



The log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed.

Parameters: μ and σ

Type: Continuous

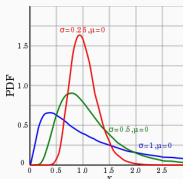
Biological scenarios: Many continuous variables that can not take negative values (e.g., weight, height).

$$\text{PDF: } \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

Range: $(0, \infty)$

Mean: $\exp\left(\mu + \frac{\sigma^2}{2}\right)$

Var: $[\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$



The gamma distribution is a continuous probability distribution that describes waiting times until a certain number of events take place. For example a gamma distribution with shape = 3 and scale = 2 is the distribution of the length of time (in years) you'd have to wait for 3 deaths to occur in a population with an average survival time of 2 years.

Parameters: shape = k and scale = θ (both >0)

Type: Continuous

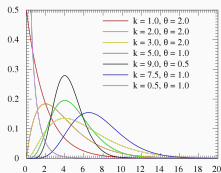
Biological scenarios: Survival time, the age distribution of cancer incidence, highly variable data where negative numbers don't make sense.

$$\text{PDF: } \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

Range: $(0, \infty)$

Mean: $k\theta$

Var: $k\theta^2$



What does all this have to do with fitting a straight line to some data you ask?

We'll get to that next lecture...

Bolker, B. M. (2008). Ecological models and data in R. Princeton University Press. Chapter 4

Hilborn R, Mangel M. The ecological detective: confronting models with data. 1997. Princeton University Press. Chapter 3