Formal Computational Skills

Lecture 9: Probability

Summary

By the end you will be able to:

- Calculate the probability of an event
- Calculate conditional probabilities with Bayes Theorem
- Use probability distributions and probability density functions to calculate probabilities
- Calculate the entropy of a system – easier than people think

Motivation: used throughout Alife/AI: classification, data mining, stochastic optimisation (ie GAs etc), evolutionary theory, fuzzy systems etc etc etc

Probability Definition

2 schools of thought:

Frequentist: probability is the number of times an event would occur if the conditions were repeated many times

Subjectivist: level of belief a rational being has that an event will occur

Not really an issue here about which school one adheres to

We will use: Probability of an event A (written as P(A)) is the number of ways A can happen divided by the total number of possible outcomes

Often calculated by empirical experiment (counting events)

Eg Suppose we throw 2 dice. Define a random variable A = sum of the 2 dice.

What is the probability that A is 10 written as: P(A=10) ?

P(A=10) = number of ways we can get 10/ number of outcomes

	1	2	3	4	5	6	
1	1, 1	1, 2	1, 3	1, 4	1, 5	1,6	1/6
2	2, 1	2, 2	2, 3	2, 4	2, 5	2,6	1/6
3	3, 1	3, 2	3, 3	3, 4	3, 5	3,6	1/6
4	4, 1	4, 2	4, 3	4, 4	4, 5	4,6	1/6
5	5, 1	5, 2	5, 3	5, 4	5, 5	5,6	1/6
6	6, 1	6, 2	6,3	6,4	6,5	6,6	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

We get 10 if we get {5,5}, {6,4} or {4,6} and there are 36 possible outcomes so:

P(A=10) = 3/36 = 1/12

If 2 events are independent, the outcome of one has no bearing on the outcome of the other (eg tossing 2 dice)

The probability of 2 independent events A and B happening is:

$$P(A,B) = P(A)\,P(B)$$

eg P(1, 3) = 1/6 x 1/6 = 1/36

B

A

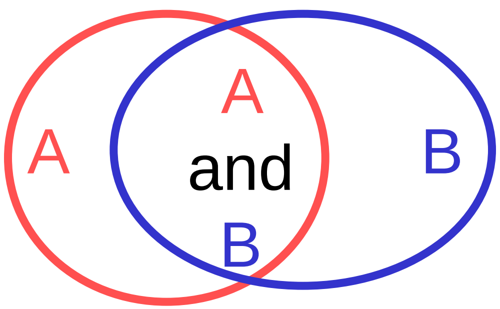
	1	2	3	4	5	6	
1	1, 1	1, 2	1, 3	1, 4	1, 5	1,6	1/6
2	2, 1	2, 2	2, 3	2, 4	2, 5	2,6	1/6
3	3, 1	3, 2	3, 3	3, 4	3, 5	3,6	1/6
4	4, 1	4, 2	4, 3	4, 4	4, 5	4,6	1/6
5	5, 1	5, 2	5, 3	5, 4	5, 5	5,6	1/6
6	6, 1	6, 2	6,3	6,4	6,5	6,6	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

The probability of A or B happening is

$$P(A \text{ or } B) = P(A) + P(B) - P(A,B)$$

If events can't happen together, they are mutually exclusive (eg P(B=1,B=2) =0) and thus get probabilities by adding probs. of each event eg get the marginal probabilities (prob. of each event separately) such as P(A=1) by adding 6 times 1/36 = 1/6

If not, must take away P(A, B) or it would be added twice:



P(2 or 4)= 1/6 + 1/6 - 1/36
= 11/36

	1	2	3	4	5	6	
1	1, 1	1, 2	1, 3	1, 4	1, 5	1,6	1/6
2	2, 1	2, 2	2, 3	2, 4	2, 5	2,6	1/6
3	3, 1	3, 2	3, 3	3, 4	3, 5	3,6	1/6
4	4, 1	4, 2	4, 3	4, 4	4, 5	4,6	1/6
5	5, 1	5, 2	5, 3	5, 4	5, 5	5,6	1/6
6	6, 1	6, 2	6,3	6,4	6,5	6,6	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

Conditional probabilities

What if we want $P(A=10)$ where A is sum of 2 dice but we have already thrown one die?

Use conditional probability: $P(A|B)$ where B represents throwing the first die

Probability of an event A happening if event B has already happened

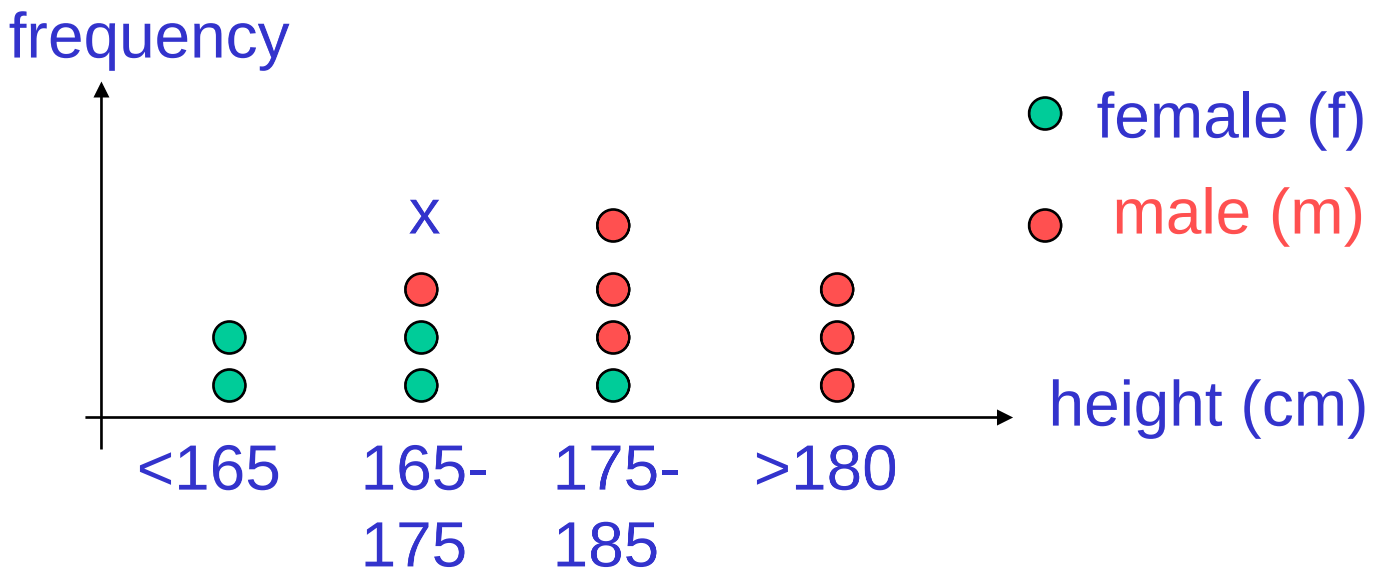
read $P(A|B)$ as probability of A given B

eg if we have already got a 5, probability that the sum is 10 is:

$$P(A=10|B=5) = 1/6$$

as we must now get a 5 with the second die

Eg: data classification. Classify person's sex from their height



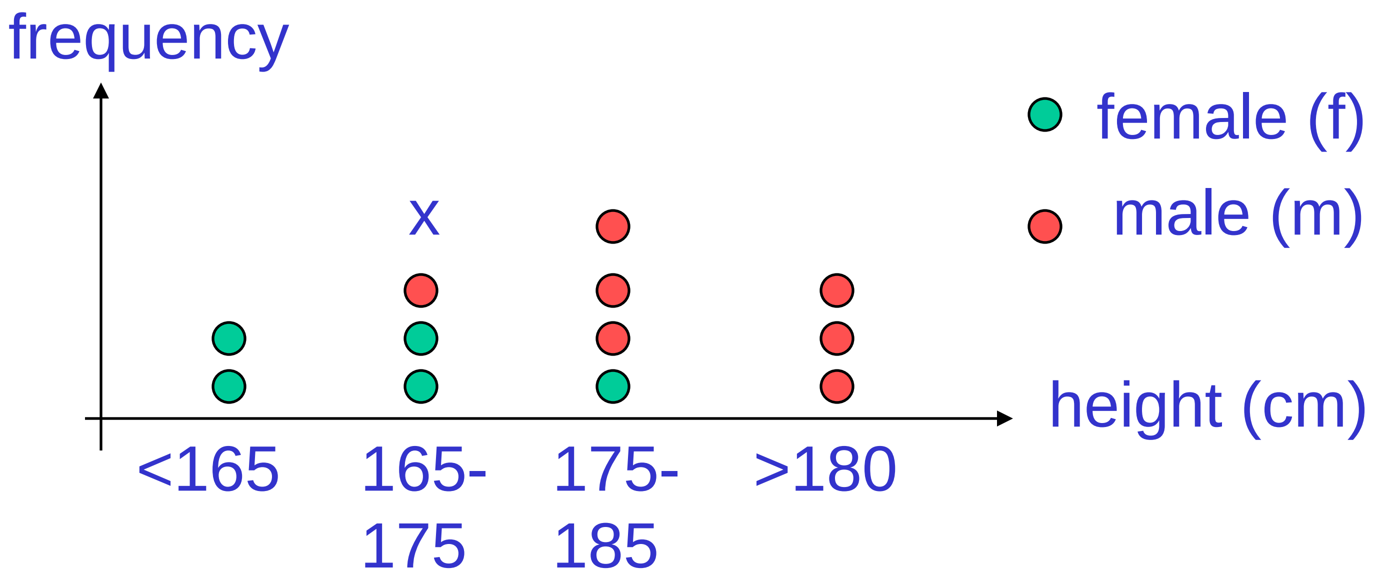
P(m)? 7 out of 12 in sample so P(m) =7/12 and P(f) = 5/12
Similarly, P(h=165-175) = 3/12

What sex is x?? if we didn't know height would go for male as from our survey this is more likely: P(m)>P(f)

However we know h = 167:
P(m|h=167) = 1/3 < p(f|h=167) so go for female

Bayes Theorem

Very (very) useful theorem: $P(A|B) = P(B|A)\,P(A)/P(B)$



eg $P(m|h=167) = P(h=167|m)\,P(m)/P(h=167)$

$= 1/7 \times 7/12 \,/\, 3/12 = 1/3$

Used since it can be easier to evaluate P(B|A) (eg here we don't need to know anything about females' heights)

Note for classification, just need whether $P(m|h)>P(f|h)$ so compare $P(h|m)\,P(m)$ with $P(h|f)\,P(f)$ as $P(h)$ just normalises the probabilities

$P(m)$ known as prior probability (or just prior) as it is the probability prior to getting any evidence/data

$P(m|h)$ known as the posterior probability

Biggest problem in Bayesian inference is estimating the priors correctly and often they are simply assumed to be equal eg $P(m)=P(f)=1/2$

Very important as they encapsulate our prior knowledge of the problem and can have a very big impact on the outcome

Often priors estimated iteratively (eg in SLAM) starting with no assumptions and incorporating evidence at each time-step

Probability Distributions

Suppose we have a 2d binary random variable X (random variables written with a capital letter) where each dimension has equal probability of being 0 or 1.

What are the different states (written as a small letter ie x)?

What are the probabilities of each state (written as P(X=x))?

B1	B2	X=(B1,B2)	P(B1,B2)
0	0	(0,0)	0.5x0.5 = 0.25
0	1	(0,1)	0.25
1	0	(1,0)	0.25
1	1	(1,1)	0.25

The 3rd and 4th columns form the probability distribution of X

Probability distribution tells us the probability of each state of X occurring

Where possible it is given as a function

EG suppose Y = B1+B2. What are the different states of Y?

B1	B2	Y=(B1+B2)	P(B1,B2)
0	0	?	0.25
0	1	?	0.25
1	0	?	0.25
1	1	?	0.25

What is the probability distribution?

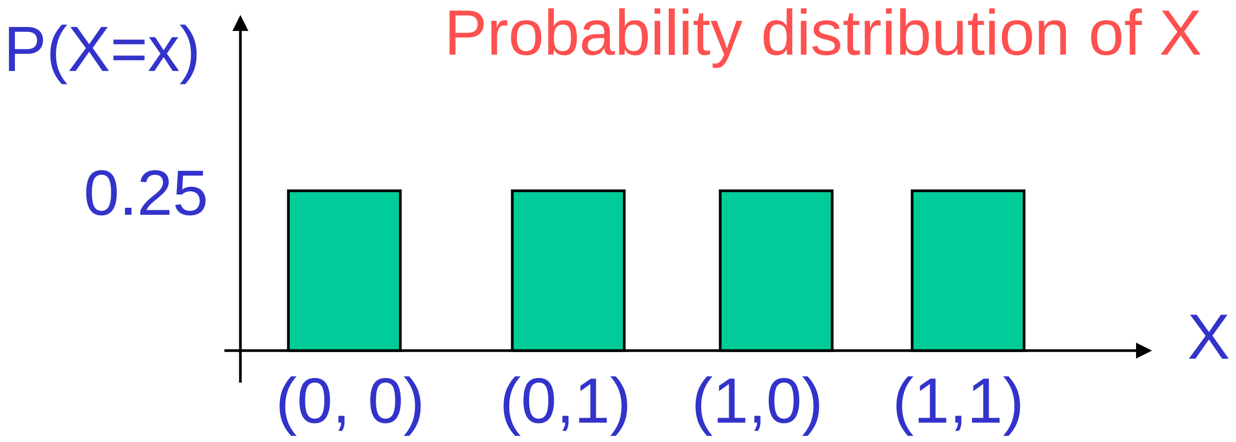
EG suppose Y = B1+B2. What are the different states?

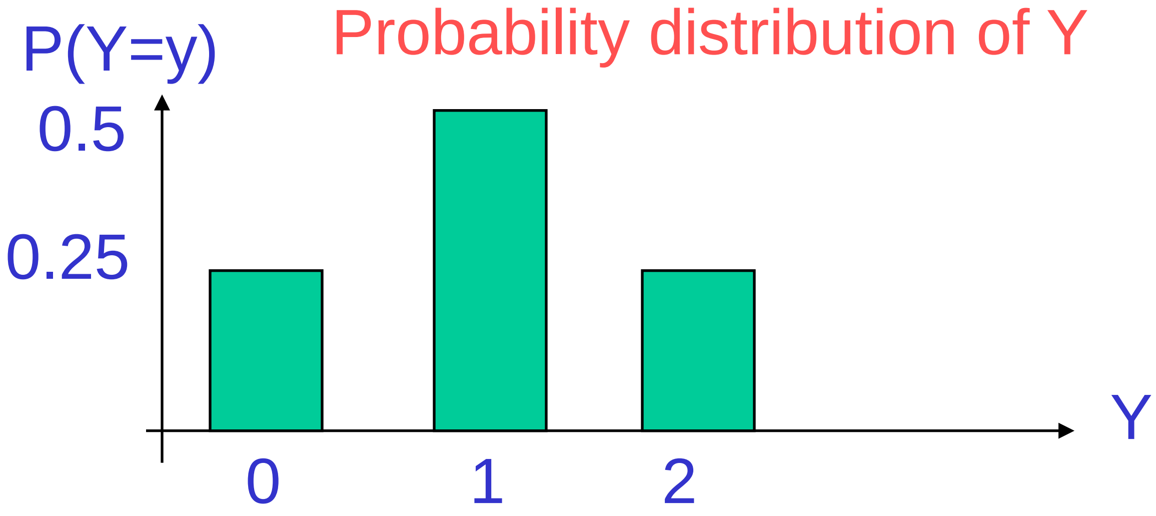
B1	B2	Y=(B1+B2)	P(B1,B2)
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	2	0.25

so the probability distribution of Y is

Y=(B1+B2)	P(Y=y)
0	0.25
1	0.5
2	0.25

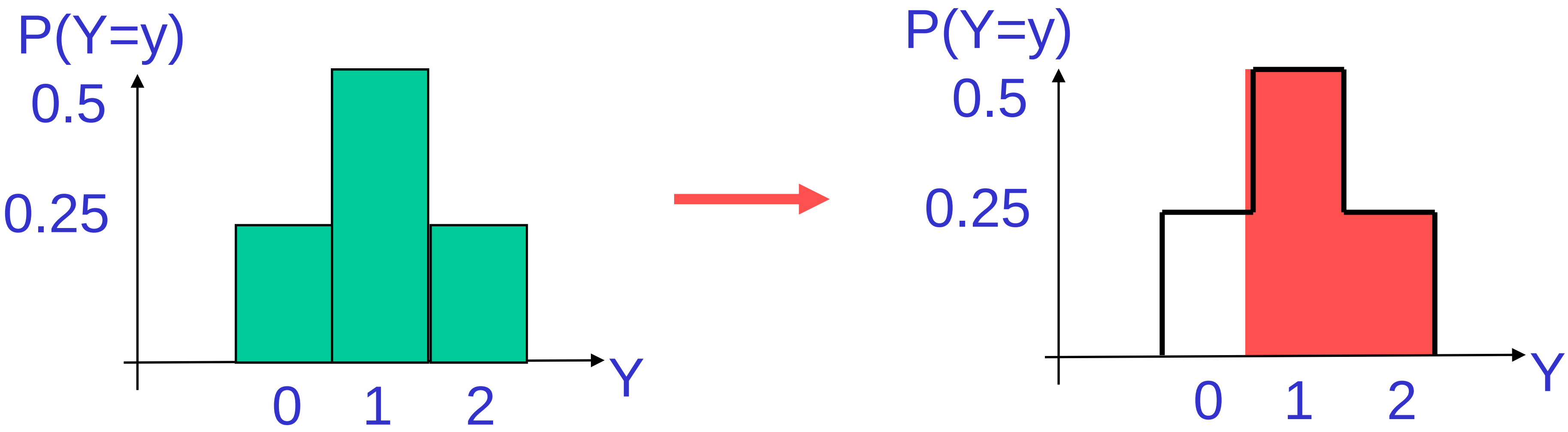
Often view probability distributions graphically:





Note that probabilities sum to 1.

Also, if we make each state (ie bar) width one, and put them together, can draw as a graph



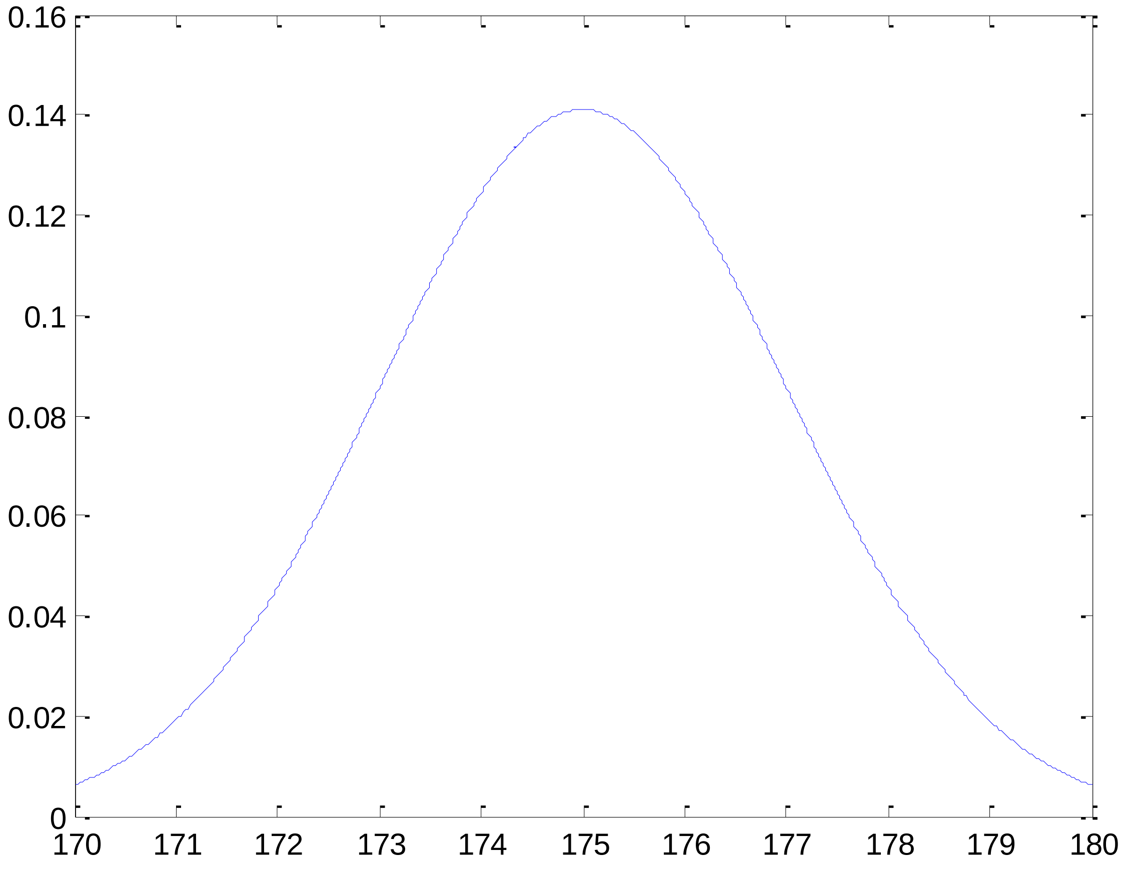
Probability of one or many states is area of bars ie area under graph

Eg P(Y = 1 or 2) = 0.75

Probability Density Functions

In the previous eg's, both X and Y are discrete random variables as the states are discrete

What about continuous variables such as people's heights??

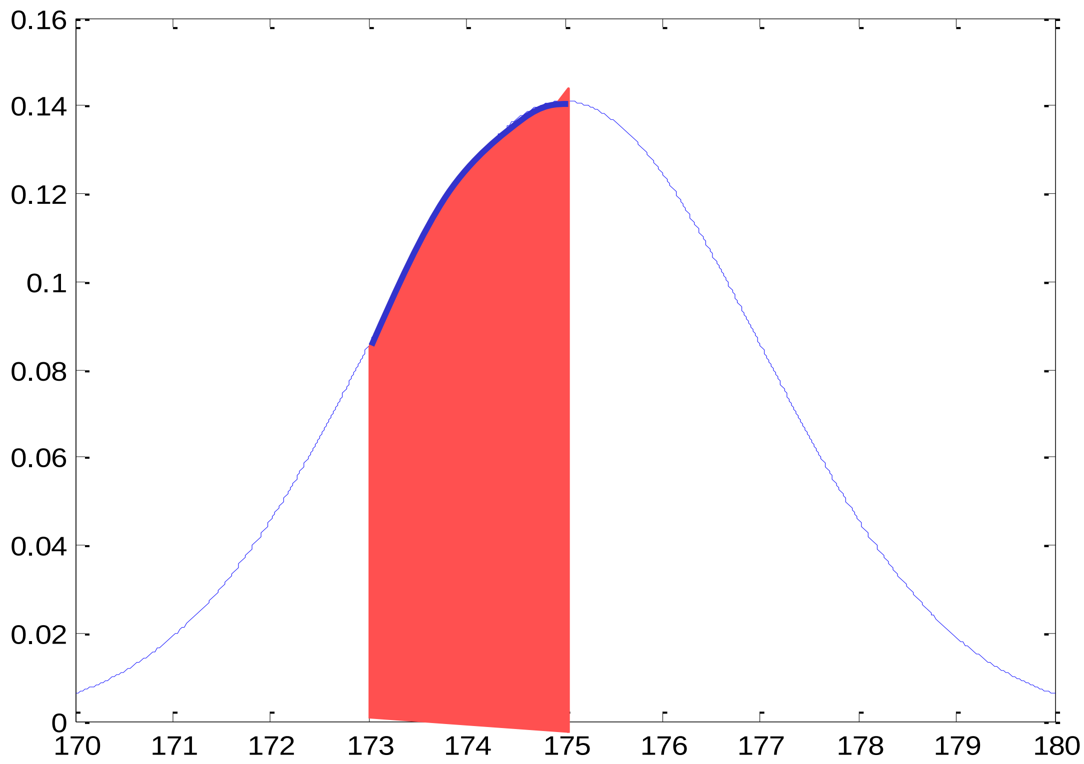


Again use a function but a continuous one

Function known as a probability density function (pdf)

Value f(x) indicates the likelihood of each height eg 175 is the commonest height

However, probability of anyone's height being exactly 175 is 0. Instead get probability height is in a range of values eg between 173 and 175



As with discrete variables area under the curve gives the probability so if pdf is f(x)

$$P(173 \leq X \leq 175) = \int_{173}^{175} f(x)dx$$

Again area under curve must sum to 1

Properties/Parameters of Distributions

Distributions and pdfs are often described/specified by parameters, commonly mean and variance

The mean is the average value and is also known as the expected value E(X) or <X>

Variance governs the spread of the data and is the square of the standard deviation

For n-dimensional data, also have the nxn covariance matrix where the ij’th element specifies the variance between the i’th and j’th dimensions

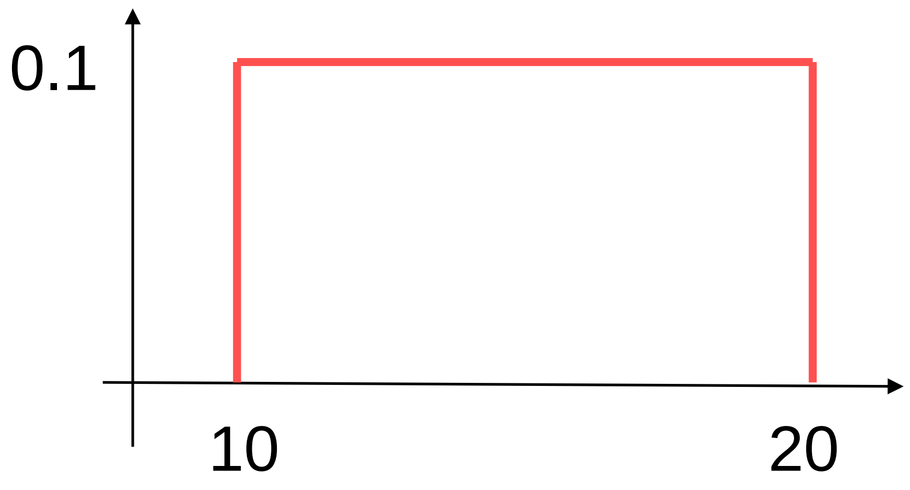
Common Distributions

Have met the 2 most common: uniform and gaussian or normal

Often used as the mutation operators in a GA, hill climbing etc

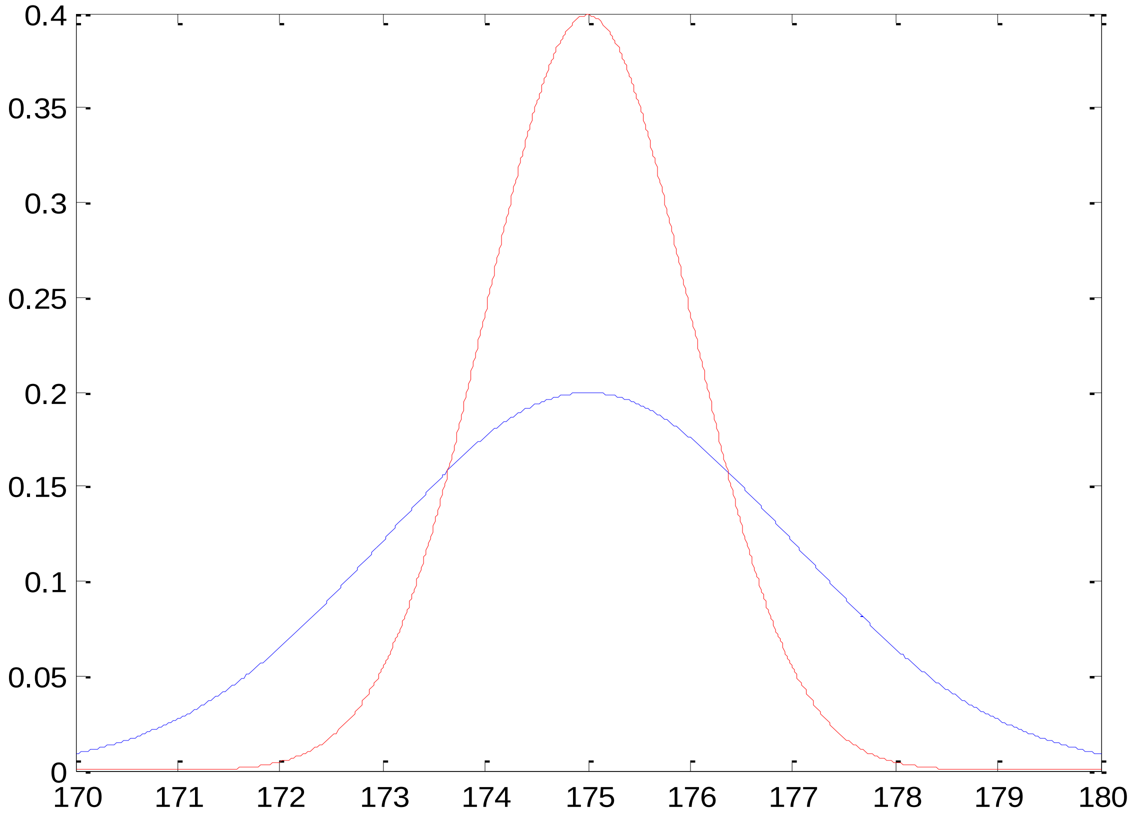
Other distributions include binomial, multinomial, poisson etc

Uniform distribution has the same probability for each point.
Thus probability is governed by the range of the data R and pdf
$f(x) = 1/R = 1/10 = 0.1$



Gaussian is governed by mean μ and variance σ^2 with pdf:

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)}$$



red has $\sigma^2 = 1$

blue has $\sigma^2 = 4$

It is centred on μ and width (and height) are governed by σ^2

Properties of the Gaussian

Gaussian may look nasty but has lots of nice properties including:

- Sum of gaussians is gaussian
- Conditional densities are gaussian
- Linear transformation is gaussian
- Easy to work with mathematically
- For a given mean and variance, entropy (see later) is maximised by gaussian distribution
- Central Limit theorem: sum/mean of m uniform random variables is gaussian

Entropy

For a discrete distribution, entropy is:

$$S(X) = -\sum_{x \in X} P(X = x) \ln(P(X = x))$$

and for continuous:

$$S(X) = -\int_{x \in X} f(x) \ln(f(x)) dx$$

Doesn't depend on values of X but how probability is spread

Entropy can be formulated in many ways but can be seen as a measure of smoothness of a distribution

Heavily used in information theory: the average amount of information needed to specify a variable's value is the entropy

If log is to the base e information measured in nats, if log is to base 2 ie

$$S(X) = -\int_{x \in X} f(x)\log_2(f(x))dx$$

info is measured in bits

Can also be formulated as degree of surprise

suppose probability of one state is 1 and all others = 0. Entropy is 0 as log(1) = 0 and 0 log (0) = 0

Intuitively makes sense as event MUST happen, so no information is passed and similarly no surprise in hearing it has happened

Conversely, if variable is entirely random, entropy is large as we must specify all variables

If X is a 1d binary variable, how many bits of information needed to specify it?

Ans =1.

What is entropy? remember that $\log_2(2^{-n}) = -n$

X	P(X=x)	$\log_2(P(X=x))$
0	$0.5 = 2^{-1}$	-1
1	$0.5 = 2^{-1}$	-1

$$S(X) = -\sum_{x \in X} P(X = x) \log_2(P(X = x))$$

So S(X)= -(0.5 (-1) + 0.5 (-1)) = 1

If X is a 2D binary variable, how many bits are needed to specify each state?

B1	B2	X=(B1,B2)	P(B1,B2)
0	0	(0,0)	$0.25 = 2^{-2}$
0	1	(0,1)	$0.25 = 2^{-2}$
1	0	(1,0)	$0.25 = 2^{-2}$
1	1	(1,1)	$0.25 = 2^{-2}$

Ans =2.

What is entropy? remember that $\log_2(2^{-n}) = -n$

$$S(X) = -\sum_{x \in X} P(X = x) \log_2(P(X = x))$$

$S(X) = -(0.25\ (-2) + 0.25\ (-2) + 0.25\ (-2) + 0.25\ (-2)\ )$
$= -4(0.25\ (-2)) = 2$

For nD binary variable, have 2^n states what is probability of each?

Ans 2^{-n}. What about entropy/ no. of bits to specify state?

$S(X) = -2^n\,(2^{-n}\,(-n)) = n$

What about Y = B1+B2. Will entropy of Y be more or less than X?

Y=(B1+B2)	P(Y=y)
0	$0.25 = 2^{-2}$
1	$0.5 = 2^{-1}$
2	$0.25 = 2^{-2}$

What is entropy?

Ans: $S(X) = -(0.25(-2) + 0.5(-1) + 0.25(-2))$

$= 1.5$

Less entropy than X as more order as more predictable

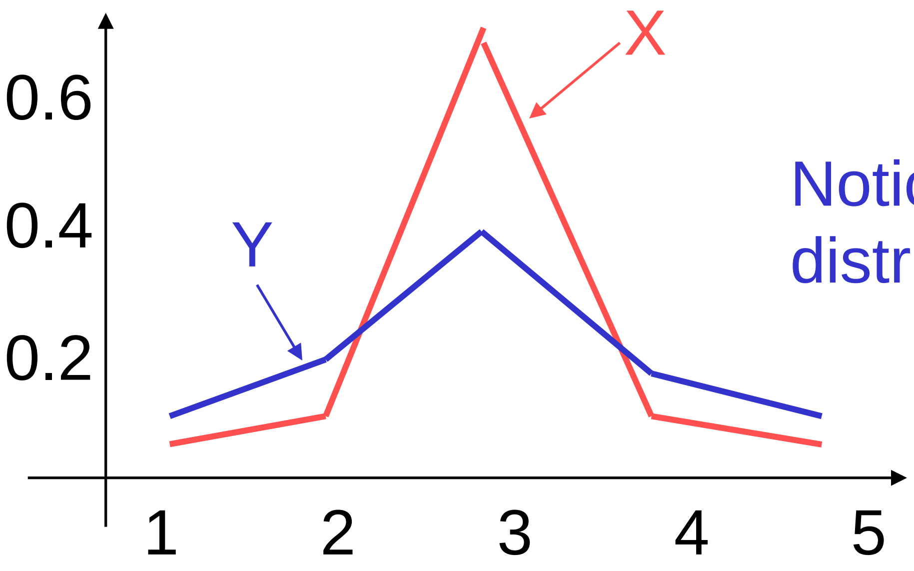
eg if you had to guess Y you would say 1: 50:50 chance of being right whereas for X 2D random binary, only 25% chance

Which of X and Y will have higher entropy?

	P(X=x)	$\log_2(P(X=x))$	P(Y=y)	$\log_2(P(Y=y))$
1	0.05	-4.3	0.1	-3.3
2	0.1	-3.3	0.2	-2.3
3	0.7	-0.5	0.4	-1.3
4	0.1	-3.3	0.2	-2.3
5	0.05	-4.3	0.1	-3.3

Entropy of X= 1.45

Entropy of Y = 2.12



Notice that the steeper distribution has less entropy

For a given mean and variance, entropy is maximised by gaussian distribution