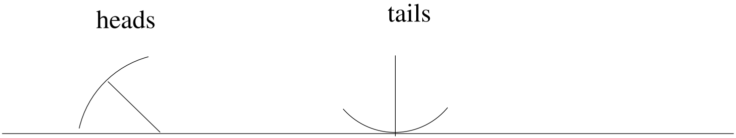
Overview

- Here we address an old criticism of Bayes nets (and quantitative methods in AI generally):
- Where do the numbers come from?
- The numbers here are probabilities and we get them from data.

Estimating probabilities

- Suppose we wish to estimate the probability θ that a certain drawing pin lands heads. We toss it 100 times and it comes up heads 35 times. What's our best guess for θ?
- If we had tossed it once, and it had come up heads, what would be our guess for θ then?



Formalising our assumptions

We have data points (drawing-pin tosses) $D = D_1, D_2, \dots D_n$ where

1. each is sampled from the same distribution
2. each is independent

Such samples are *independent and identically distributed* or *iid*.

The likelihood function

- The distribution here is parameterised by a single parameter θ, which defines a probability $P(D|\theta)$
- (In other cases, θ will be a vector of parameters.)
- For a data set D, we define the *likelihood function*:

$$L(D|\theta) = P(D|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

What's the likelihood of the sequence h, t, t, t, h, h?

Sufficient statistics

- The likelihood only depends on the number of heads N_h and tails N_t, not, say, the order in which they occurred
- N_h and N_t are therefore *sufficient statistics*
- A *sufficient statistic* is a function of the data that summarises the relevant information for computing the likelihood.

Formally, $s(D)$ is a sufficient statistic if, for any two datasets, D and D':

$$s(D) = s(D') \Rightarrow L(D|\theta) = L(D'|\theta)$$

Maximum likelihood estimation

- The *maximum likelihood estimation* (MLE) principle tells us to choose that value of θ which maxmises the likelihood (for the observed data).
- This value (often denoted $\hat{\theta}$) is (apparently) the best **estimate** for θ
- This is the value of θ which makes the data as likely as possible.
- What's $\hat{\theta}$ for our drawing pin?

Classical Statistics

- MLE is an example of *Classical* or *non-Bayesian* statistical inference
- θ is treated as an objectively fixed, but unknown value
- Therefore it does not make sense to talk of eg the probability that θ lies in the interval $(0.3, 0.4)$ the unknown θ is either definitely in that interval or not, so we can't talk of probability in this context.
- The data, on the other hand, does have a probability - why is this OK?

Problems with the Classical approach

- Rather than give us the most likely value for θ given the data …
- …MLE gives us that $\hat{\theta}$ such that the data is as likely as possible
- This is basically the wrong way round,
- but to be able to talk about $P(\theta|D)$ we have to have a probability distribution over θ

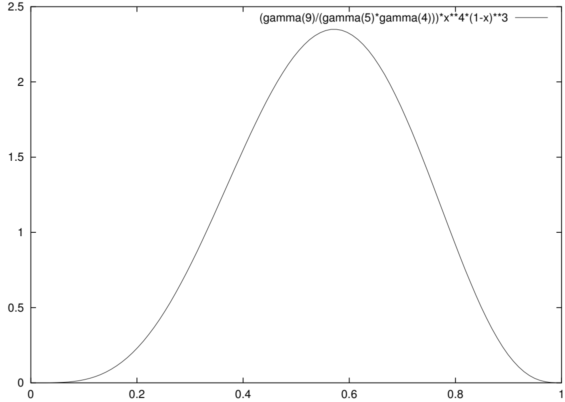
The Bayesian paradigm

- The Bayesian approach to statistics permits probability to represent *subjective uncertainty*
- It was a minority view until quite recently, since subjectivity was seen as unscientific
- More popular now partly because there are better tools available.
- For example, the BUGS system (Bayesian inference using Gibbs sampling)

Bayesian estimation of probabilities - prior

- We express our uncertainty about the true value of θ by placing a *prior distribution* over possible values of θ
- This distribution is defined (somehow!) **prior** to the collection of data
- Since there are uncountably infinitely many values of θ the distribution is represented by a *probability density function* - here's one:

Prior distribution over θ



This is a graph of $f(\theta) = \mathrm{Beta}(\theta|5,4) = \frac{\Gamma(9)}{\Gamma(5)\Gamma(4)}\theta^4(1-\theta)^3$

More on the Beta distribution later

Some points about density functions

- $\mathrm{Beta}(\theta|5,4)(x)$ does **not** give our prior probability that $\theta = x$
- To get probabilities out of density functions we integrate
- To get the prior probability that $\theta \in (0.3, 0.4)$, we compute:

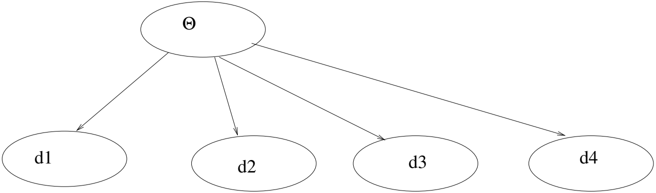
$$\int_{0.3}^{0.4} \mathrm{Beta}(\theta|5,4)(x)dx$$

- Unfortunately, there is no closed form for this integral - a fact which upset Rev Bayes considerably

Connecting prior and evidence

- The hallmark of Bayesian analysis is that everything is treated as a random variable - both the unknown parameter θ and the data D
- θ is of course never observed
- D - the data - is always observed (let us assume that for now anyway).
- Since everything is a random variable, we can use a Bayesian network to represent the joint distribution over (θ, D).

Bayes net



Instead of a table for the distribution over θ we have the density function

The conditional probabilities (which can not be represented by CPTs) are all identical,

$$\forall i : P(D_i = h|\theta = x) = x$$

Or, for short, $P(D_i = h|\theta) = \theta$

Beta distributions

A beta distribution is determined by two parameters, usually denoted α and β:

$$\mathrm{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Its *mean* value is $\frac{\alpha}{\alpha+\beta}$. If $\mathrm{Beta}(\theta|\alpha,\beta)$ represents our current beliefs about the likely whereabouts of θ, then this mean value is a good estimate for θ.
Its *mode* is at: $\frac{\alpha-1}{\alpha+\beta-2}$

Why the Beta distributions for estimating probabilities

- Suppose $\mathrm{Beta}(\theta|\alpha, \beta)$ represents our beliefs about θ, where θ is the (true) probability that the drawing pin lands heads.
- Suppose we toss the drawing pin and it lands heads.
- Our new *posterior distribution* is simply $\mathrm{Beta}(\theta|\alpha + 1, \beta)$!
- Our new mean is just $\frac{\alpha+1}{\alpha+1+\beta}$
- If it had landed tails, we would have $\mathrm{Beta}(\theta|\alpha, \beta + 1)$

Experience

- Following Netica, we can call $\alpha + \beta$ *experience*. This increases by one each time we observe a drawing pin toss (or equivalent). It is also called the *effective sample size*, since it reflects how many pieces of data you have observed (or pretend to have observed).
- The larger it is the more pointy the beta distribution is—which makes sense.
- The mean ($\frac{\alpha}{\alpha+\beta}$) and experience ($\alpha + \beta$) determine a beta distribution.

Estimating conditional probabilities

- Estimating conditional probabilities is not really different from estimating any other sort of probability.
- To estimate, say $P(A = true|B = false)$ from a series of observations of A and B just
    1. Ignore any cases where $B = true$
    2. Where $B = false$, use the observed values of A to update as above.

Multinomial probabilities

- If we wanted to estimate the 6 probabilities associated with a die throw, the beta distribution would be inappropriate.
- Using the *Dirichlet distribution*, we can estimate all 6 probabilities simultaneously from a sequence of die throws.
- Beta distribution is just a Dirichlet distribution where $k = 2$

Dirichlet distribution: the grisly details

$$\mathrm{Dirichlet}(\theta_1, \ldots, \theta_n | \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Mean value for θ_i is $\frac{\alpha_i}{\alpha_0}$ where $\alpha_0 = \sum_{j=1}^{k} \alpha_j$ is the "experience" as before.

Think of the α_i as counts.

Bayesian learning with Netica

- Each conditional probability in Netica has a (hidden) Dirichlet distribution associated with it.
- The conditional probability you see is the mean of this distribution.
- The initial experience is set to 1.

Bayesian learning with Netica

- Netica assumes that the θ for each conditional probability are independent (this is called *parameter independence*)
- These θs are random variables, but are **not** represented as nodes in Netica networks (they are in BUGS).