The Exciting Guide To Probability Distributions – Part 2

Jamie Frost – v1.1

Contents

- Part 2
  - A revisit of the multinomial distribution
  - The Dirichlet Distribution
  - The Beta Distribution
  - Conjugate Priors
  - The Gamma Distribution

BACK TO THE FUTURE MULTINOMIAL DISTRIBUTION

- We saw in the last part that the multinomial distribution was over counts of outcomes, given the probability of each outcome and the total number of outcomes.

$$f(x_1, \ldots, x_k \mid n, p_1, \ldots, p_k) = [n! / \prod x_i!] \prod p_i^{x_i}$$

The count of each outcome.

The probability of each outcome.

- That's all smashing, but suppose we wanted to know **the reverse**, i.e. the probability that the distribution underlying our random variable has outcome probabilities of $p_1, \ldots, p_k$, given that we observed each outcome $x_1, \ldots, x_k$ times.
- In other words, we are considering all the possible probability distributions $(p_1, \ldots, p_k)$ that could have generated these counts, rather than all the possible counts given a fixed distribution.

BACK TO THE FUTURE MULTINOMIAL DISTRIBUTION

- Initial attempt at a probability mass function: Just swap the domain and the parameters:

The RHS is exactly the same.

$$f(p_1, \ldots, p_k \mid n, x_1, \ldots, x_k) = [n! / \prod x_i!] \prod p_i^{x_i}$$

- Notational convention is that we define the support as a vector **x**, so let's relabel **p** as **x**, and the counts **x** as **α**...

$$f(x_1, \ldots, x_k \mid n, \alpha_1, \ldots, \alpha_k) = [n! / \prod \alpha_i!] \prod x_i^{\alpha_i}$$

- We can define n just as the sum of the counts:

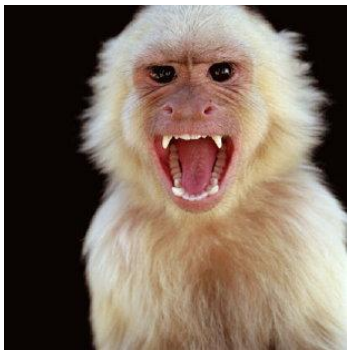
$$f(x_1, \ldots, x_k \mid \alpha_1, \ldots, \alpha_k) = [(\sum \alpha_i)! / \prod \alpha_i!] \prod x_i^{\alpha_i}$$

BACK TO THE FUTURE MULTINOMIAL DISTRIBUTION

- But wait, we're not quite there yet. We know that probabilities have to sum to 1, so we need to restrict the domain we can draw from:

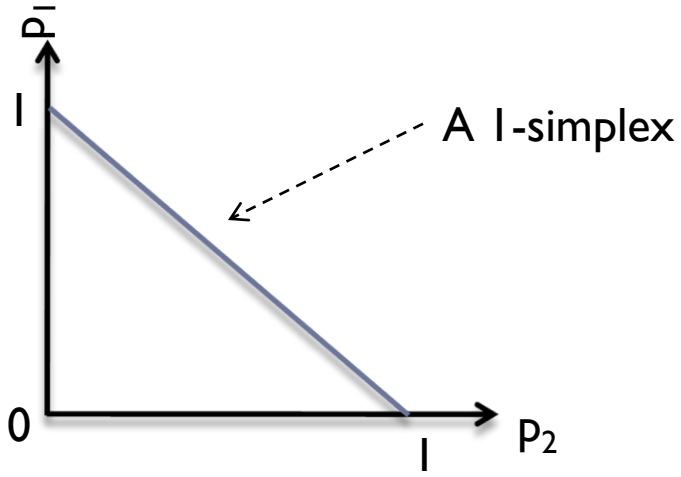
$$\text{s.t. } \sum x_i = 1$$

- We say that our distribution is defined **over the simplex**. What the rabid monkey does that mean?!



Simplex

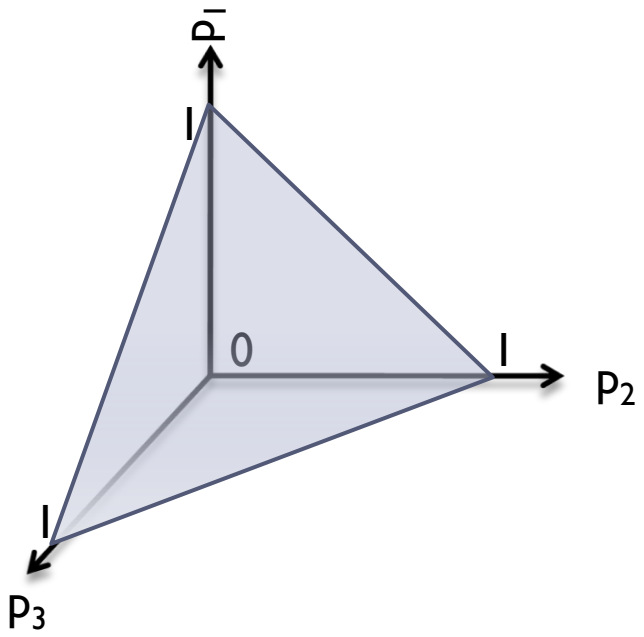
- In the 2D case where we have 2 outcomes, if we plotted p_1 and p_2 on a graph we'd get this:



- i.e. Our choice of p_1 and p_2 are not unbounded but restricted to the line-segment satisfying $p_1 + p_2 = 1$, an example of a **1-simplex**.

Simplex

- In the 3D case, we'd have a 2-simplex space, which is a triangle:

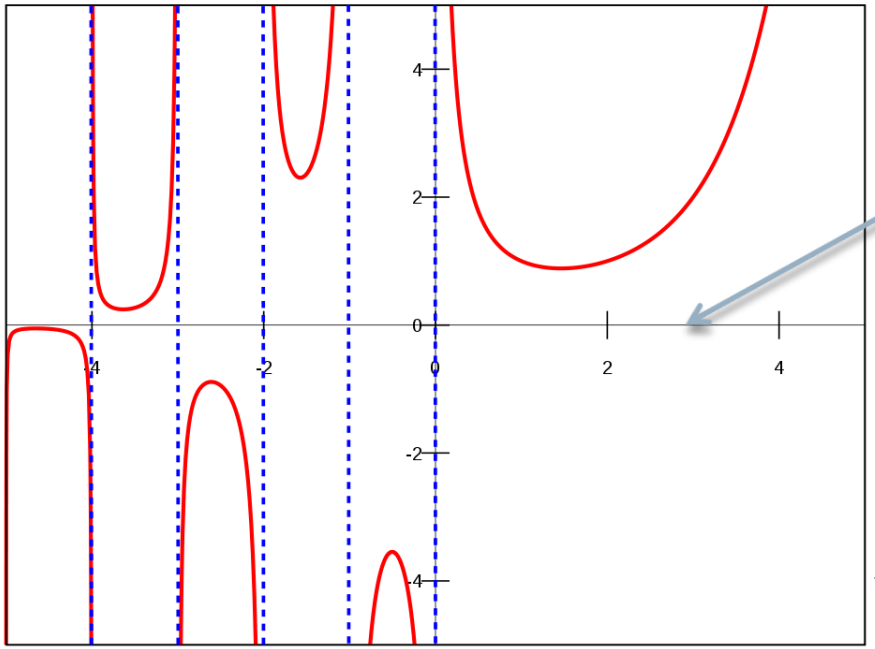


Geometrically speaking, an n-simplex is an n-dimensional polytope (i.e. a polygon/polyhedron generalised to any number of dimensions) with n+1 vertices. So a 2-simplex is a triangle, a 3-simplex is a tetrahedron, etc.

- In general for *n* probabilities our probability space is on a *n-1* simplex (we lose a degree of freedom simply because one probability can be determined by deducting all the others from 1).

The Gamma Function

- It would be nice to define our counts over real numbers. But we have $\alpha_i!$ in our equation, and the factorial function is only defined over positive integers.
- The gamma function Γ is the same as the factorial function for positive integers (if we deduct 1), but can take any real number, including negative ones!



$\Gamma(3) = (3-1)! = 2$

Dirichlet Distribution

- Putting all this together, we get:

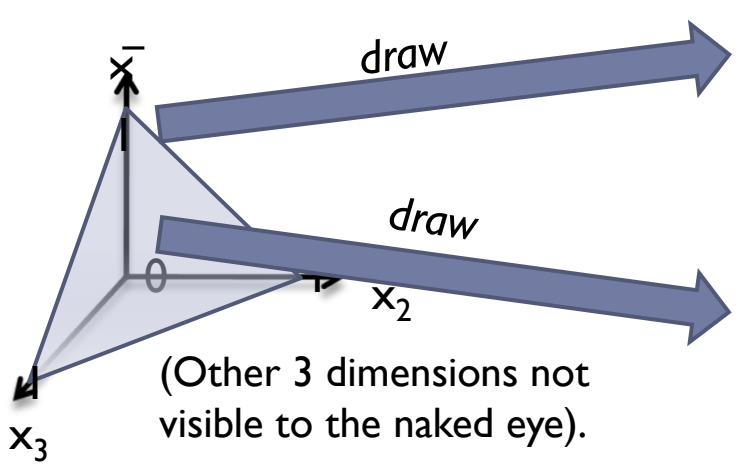
$$f(x_1, \ldots, x_k \mid \alpha_1, \ldots, \alpha_k) = [\Gamma(\sum \alpha_i) / \prod \Gamma(\alpha_i)] \prod x_i^{\alpha i - 1}$$
$$\text{s.t. } x_i > 0, \sum x_i = 1 \text{ and } \alpha_i > 0$$

?

- This is the probability density function for the Dirichlet Distribution!
- Our α_i values are known as **pseudocounts** or **concentration parameters** (we'll see what we mean by 'pseudo' later).
- It's actually cleaner to say **each event was observed $\alpha_i - 1$ times**, because we need to add 1 to any expression within a Γ so that it has the same behaviour as the factorial function.

Dirichlet Distribution

- Let's see an example. Suppose we have a dice with 6 possible outcomes, and we observed each outcome (1,2,3,4,5,6) 9 times. Thus $\alpha_i = 10 \quad \forall i$.
- Then each draw from the Dirichlet (from a 5-simplex) gives us a set of probabilities for each of the outcomes, i.e. a distribution for a single throw.

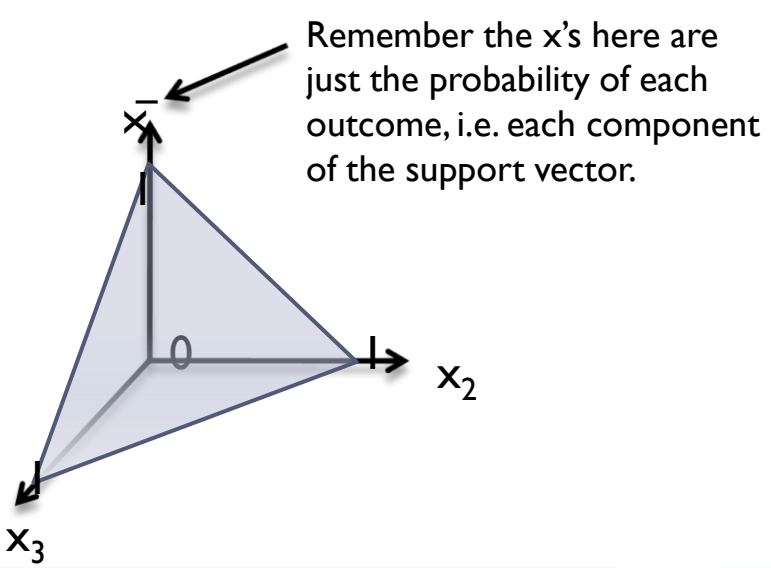


x = (1,0,0,0,0,0). i.e. A stupidly unfair dice.
Remember the p.d.f. for the Dirichlet gives us the probability that the dice actually has this distribution.

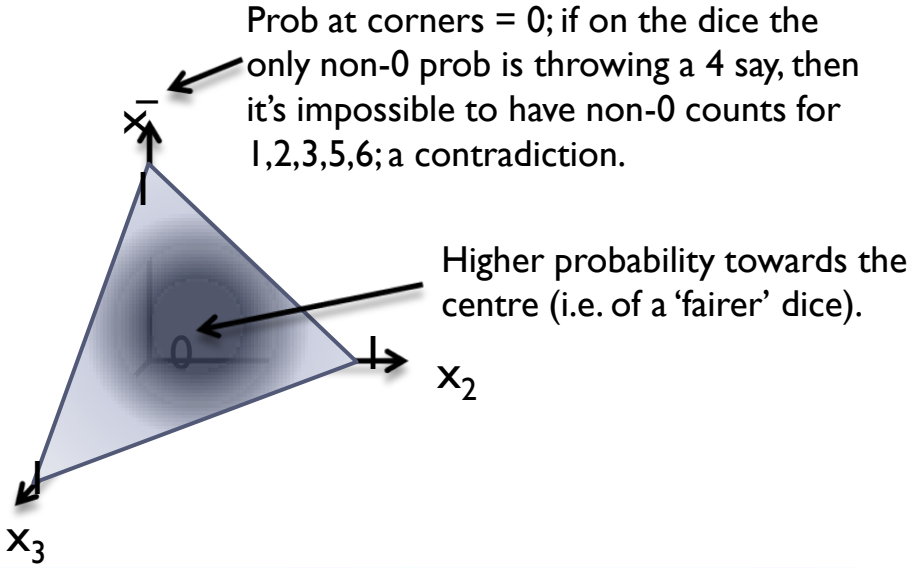
x = (1/6,1/6,1/6,1/6,1/6,1/6). i.e. A fair dice.
We've chosen a point from the middle of the simplex.

Dirichlet Distribution

- So what happens to our probability density function as we change the concentration parameters?



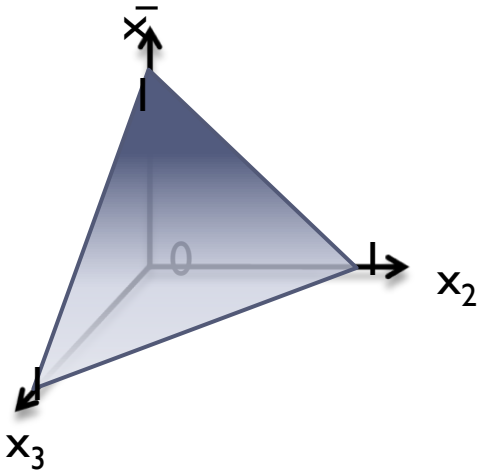
When our counts are 0 (therefore $\alpha_i = 1 \; \forall i$), we haven't made any observations, so all possible distributions are equally likely! (i.e. our Dirichlet has uniform distribution)



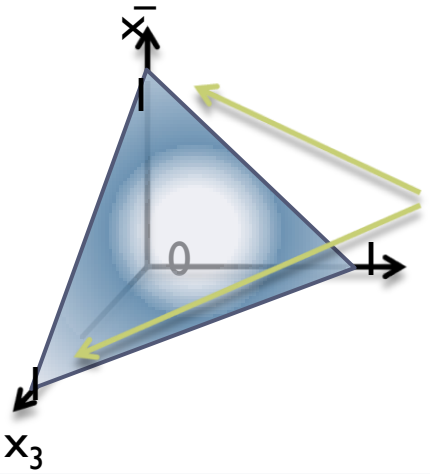
When our counts are the same but >0 (say $\alpha_i = c+1 \; \forall i$), we expect the dice to be fair, but it's still possible for unfair dice to give identical counts on multiple throws. Therefore we have a higher probability towards the centre of the simplex.

Dirichlet Distribution

- So what happens to our probability density function as we change the concentration parameters α_i ?



When one count is greater than the others, then we favour probability distributions that give a higher probability to this outcome.



Higher probability towards the corners

When the α_i are less than 1, then conceptually it's no longer helpful to think about actual counts, because they are less than 0! But the distribution just happens to have the nice property that distributions are favoured where one of the outcomes is dominant (but you don't care which).

Dirichlet

- **So just to recap...**
  - The Dirichlet Distribution gives the probability of a particular distribution given we have observed some counts of each outcome (having done repeated trials).
  - i.e. We have a distribution over distributions!
  - x_i is the probability of each outcome.
  - $\alpha_i - 1$ is the count of each outcome (and $\alpha_i > 0$).
  - It only makes sense to draw from the Dirichlet over a **simplex**, since this ensures what we draw is a valid distribution.
  - It's possible to define the Dirichlet Distribution for **fractional** and **negative** counts – that's why we call them '**pseudo counts**'.
  - Conceptually, the Dirichlet Distribution is the reverse of the Multinomial. We'll see why that's useful later...

Beta Distribution

- Just a special case of the Dirichlet Distribution, where we have just 2 outcomes.
- We can think of it as the **reverse of the Binomial Distribution** – given a number of successes out of n trials, what is the probability that our success probability p has a particular value?
- e.g. Given we flipped a coin 100 times and got 50 heads, what is the probability that p(H) = 0.5?*

* But remember we're going to have a **probability density function** here because our support (i.e. the prob of heads) is a continuous variable.

Beta Distribution

- If our concentration parameters are α and β, then our probability density function is:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

This is just a simplification from our Dirichlet p.d.f. before.

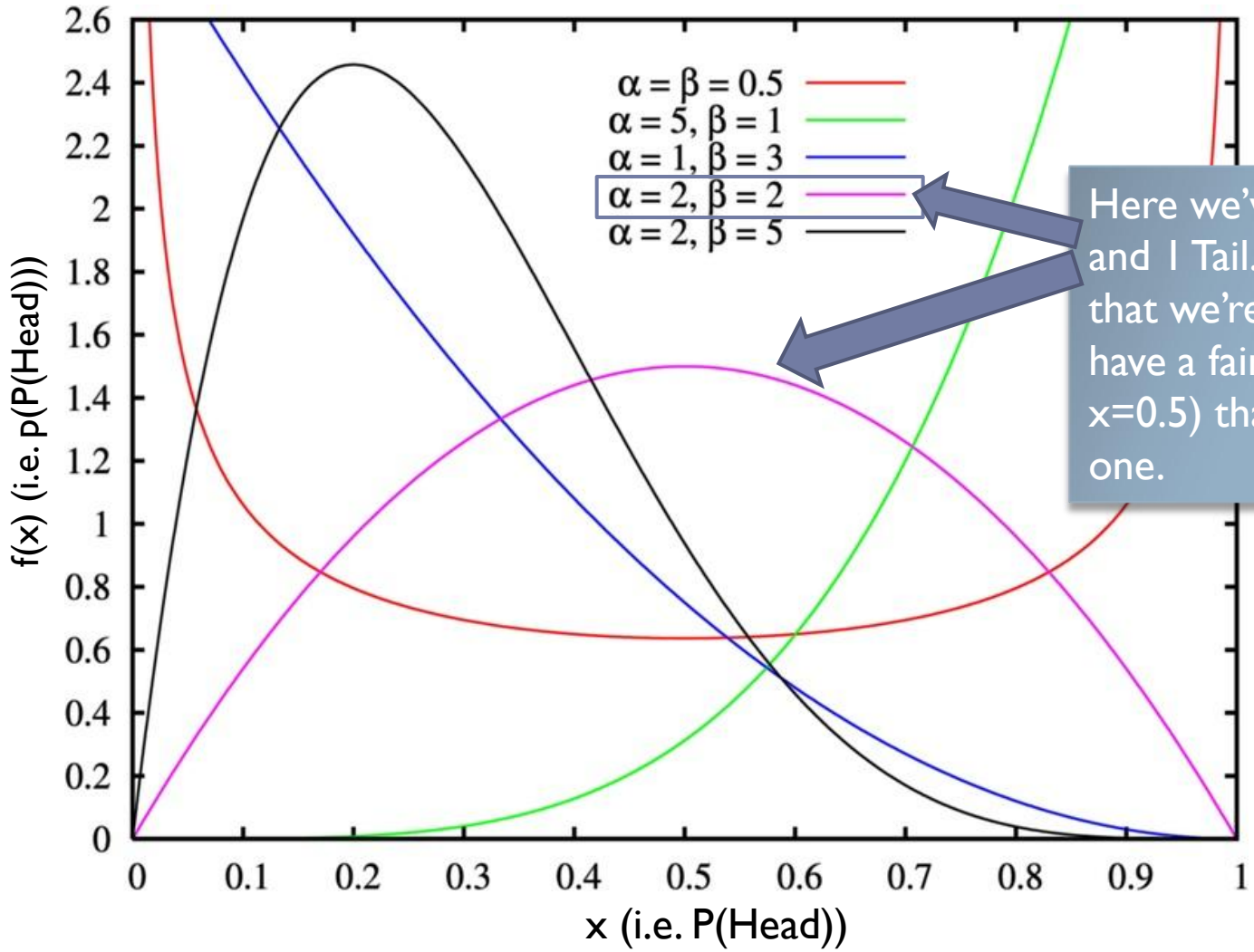
B(α,β) is known as the **beta function**.

x is a single parameter for the prob of success.*

* I hope you're not too confused by the fact that we have a **probability of a probability**. It's better to think of the probability of success as 'just a value'; a value which can vary according to some other distribution.

Beta Distribution

Sanity check: Don't be upset that our y axis goes above 1. Remember a probability density function can give a value >1; we only require that the area under the curve is = 1.



Here we've seen 1 Head and 1 Tail. It makes sense that we're more likely to have a fair coin (when x=0.5) than an unfair one.

Beta Distribution

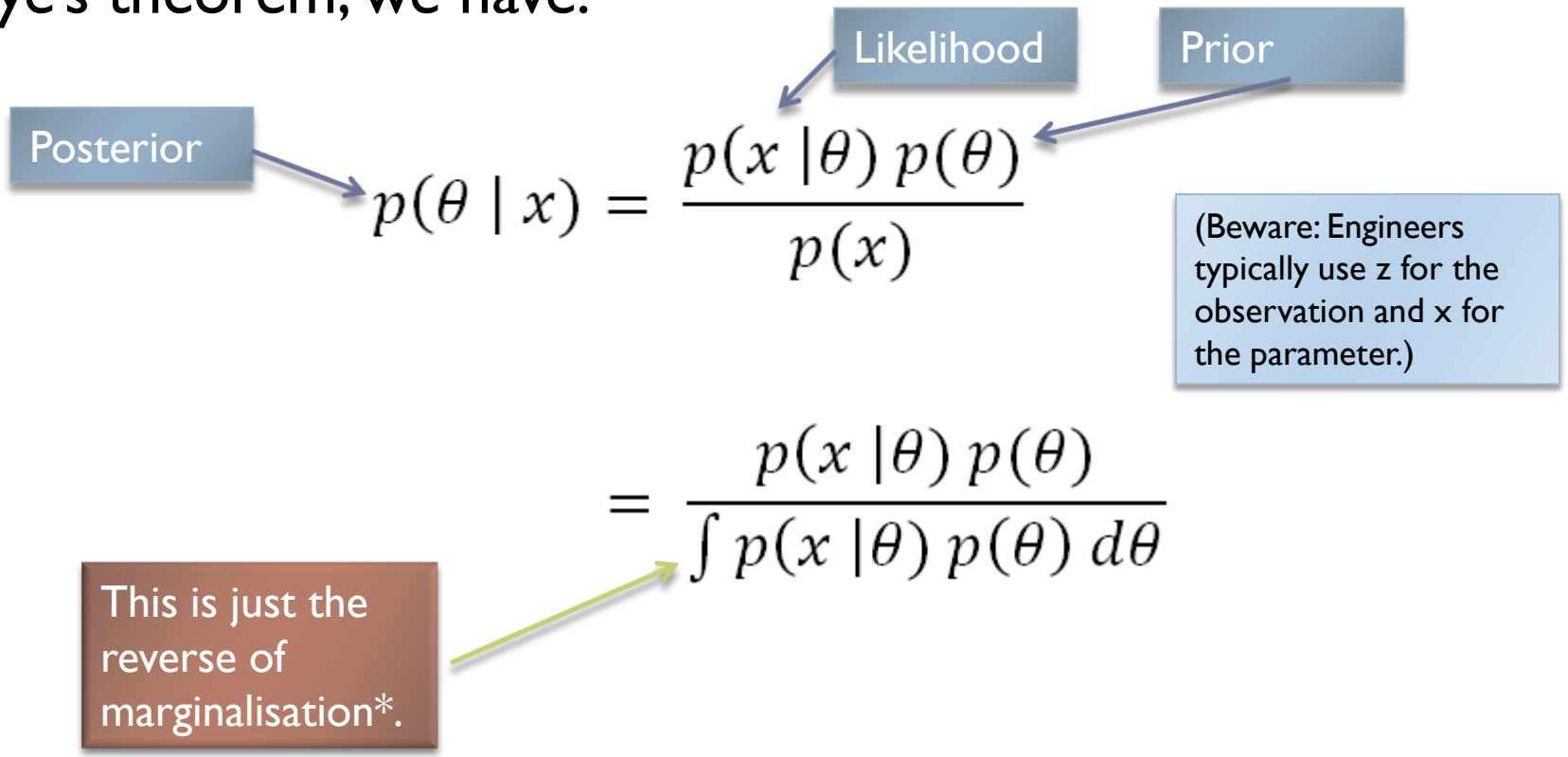
- $E[X] = \alpha / (\alpha + \beta)$ *
- We can informally see this from the Binomial Distribution. In the BD, $\alpha = np$ (where $\alpha = E[\#heads]$). Since $n = \alpha + \beta$, $E[Pr(Heads)] = p = (\alpha + \beta) / \alpha$.

- There's one last distribution called the **Gamma Distribution,** which can be conceptually thought of as the reverse of the **Poisson Distribution**, but let's have a look at conjugate priors first...

•You might be wondering why it's not $\alpha\text{-}1 / (\alpha\text{-}1 + \beta\text{-}1)$. That actually gives the **mode**, i.e. The outcome with the highest probability (or equivalently in a sample, the outcome with the highest count). If we had say 70 heads and 30 tails, then we expect the **most likely** (but NOT the 'mean' average) value of p(Heads) to be 70/100 = 0.7.
•This is quite interesting. If we have 1 head and 0 tails for example, the 'most likely' p(Heads) is obviously 1, but the (mean) average is actually 0.66. The graph in this case is linear, because trivially the probability density function of getting a head in a single throw (i.e. our stated counts) is proportional to our p(Head) value in the support.

Conjugate Priors

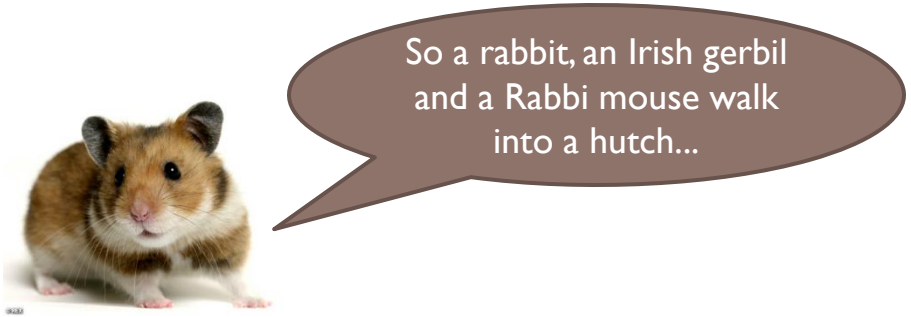
- Consider that we have some parameter θ that governs some model, and we make some observation x. From Baye's theorem, we have:



* Marginalisation just says we can get p(x) from p(x,θ) by summing over possible θ. Oh, and remember that p(x, θ) = p(x| θ)p(θ).

Conjugate Priors

- **Calculating that integral** might well be as hard as getting a hamster to do stand-up comedy...



- Suppose that we **already know** our distribution for the **likelihood** $p(x \mid \theta)$.
- It's possible to make a sensible choice for the **prior** such that our integral eventually disappears, plus **the posterior distribution takes the same form as the prior**. Now that's magic...

Conjugate Priors

- Let's take an example. Suppose we have some single parameter θ which is the probability of heads.
- Suppose our observation is the number of heads k in n flips. Then our likelihood function is just the Binomial Distribution.

$$f(k \mid n, p = \theta) = \binom{n}{k} p^k (1-p)^{n-k}$$

- But 'n' was really something that was observed (i.e. we want it on the LHS of the |), so let's use s = k (successes) and f = n-k (failures):

$$f(s, f \mid p = \theta) = \binom{s+f}{s} p^s (1-p)^f$$

Remember, it's a likelihood function because its the probability of some observation (i.e. number of heads) given some underlying parameter (our θ value).

Conjugate Priors

- So we have our given likelihood model. What can we choose for the prior, p(θ)?
- Let's try using the **beta function**. Recall our p.d.f:

(Recall the Beta function B(α, β) is the inverse of this thing on the front).

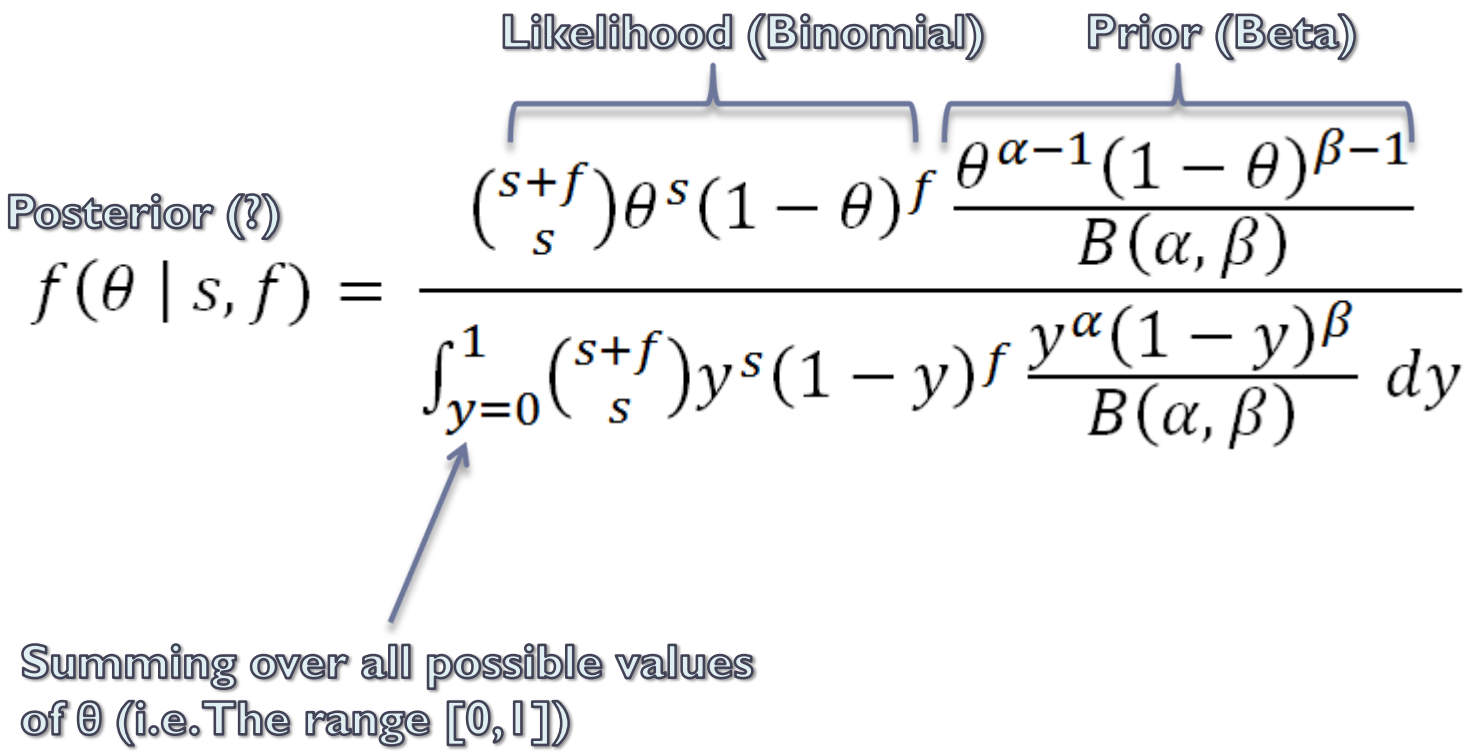
$$f(x = \theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- Remember that α-1 and β-1 were counts of heads and tails. We can choose them to reflect prior belief of what generated θ*. If we haven't the foggiest, we can just choose α = 1 and β = 1, which gives a uniform distribution (since with no counts, it's impossible to tell what the prob of heads is).
- α and β are called **hyperparameters**, to distinguish them from the underlying parameters of the model (i.e. θ).

*Important point: By counts α-1 and β-1 we don't mean the counts from the observation, we mean a **prior** belief of what counts generated θ. Otherwise it would be the posterior and not the prior!

Conjugate Priors

- Now here's the clever bit... What happens when we combine our likelihood and prior to get the posterior?



Conjugate Priors

$$f(\theta \mid s, f) = \frac{\binom{s+f}{s}\theta^{s}(1-\theta)^{f}\,\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\int_{y=0}^{1}\binom{s+f}{s}y^{s}(1-y)^{f}\,\frac{y^{\alpha}(1-y)^{\beta}}{B(\alpha,\beta)}\,dy}$$

Add the powers.

$$= \frac{\binom{s+f}{s}\,\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha,\beta)}{\int_{y=0}^{1}\binom{s+f}{s}\,y^{s+\alpha-1}(1-y)^{f+\beta-1}/B(\alpha,\beta)\,dy}$$

$$= \frac{\binom{s+f}{s}\,\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha,\beta)}{\binom{s+f}{s}\frac{B(s+\alpha,f+\beta)}{B(\alpha,\beta)}\int_{y=0}^{1}\frac{y^{s+\alpha-1}(1-y)^{f+\beta-1}}{B(s+\alpha,f+\beta)}\,dy}$$

Take the constants out of the integral + introduce $B(s+\alpha,f+\beta)$ term (such that it cancels itself out)

Conjugate Priors

$$= \frac{\binom{s+f}{s}\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha,\beta)}{\binom{s+f}{s}\frac{B(s+\alpha,f+\beta)}{B(\alpha,\beta)}1}$$

$$= \frac{\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}}{B(s+\alpha,f+\beta)}$$

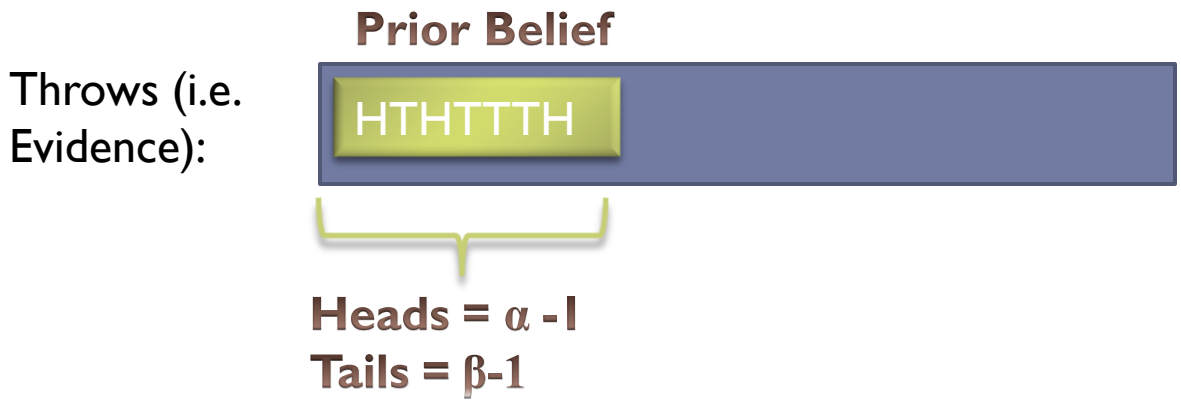
$$= Beta(s+\alpha,f+\beta)$$

The integral just becomes 1, we had the p.d.f. of a beta distribution – summing/integrating the p.d.f over all values of the support (for any valid distribution) obviously gives 1*

*If that doesn't seem obvious, then remember that for a discrete distribution, if we add up all the probabilities of each of the outcomes, they add up to 1. It's the same thing here, except that with continuous distributions we **integrate** to get 1.

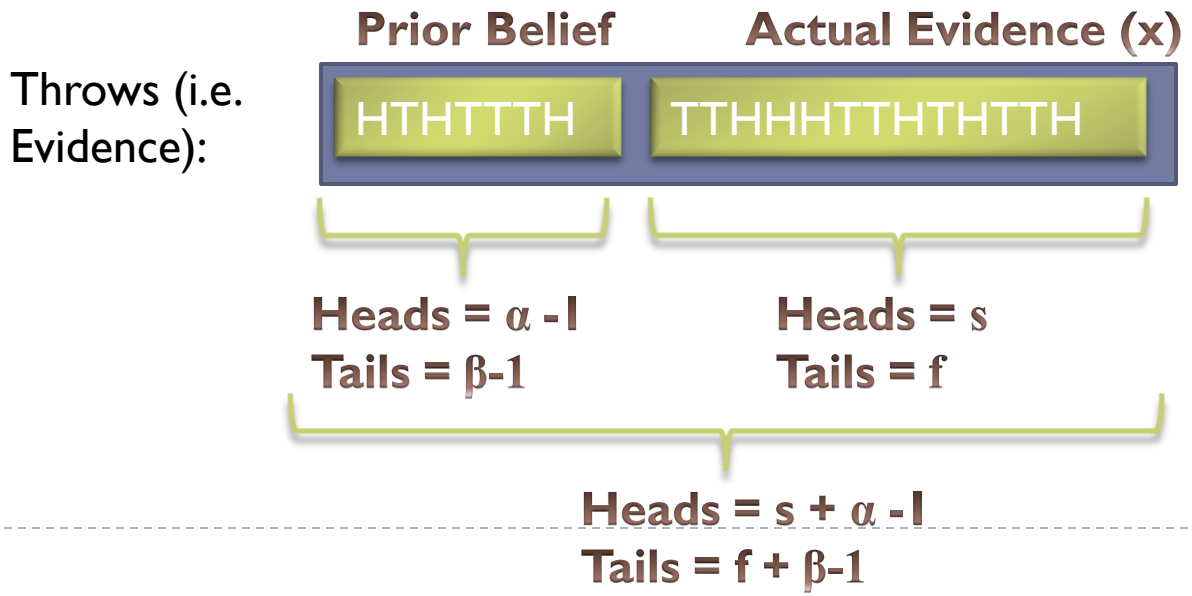
Conjugate Priors

- Is it surprising that we ended up with Beta(s+α-1,f+β-1) as the distribution for the posterior?
- Not really... For the prior we had some 'prior' evidence/belief of the counts.



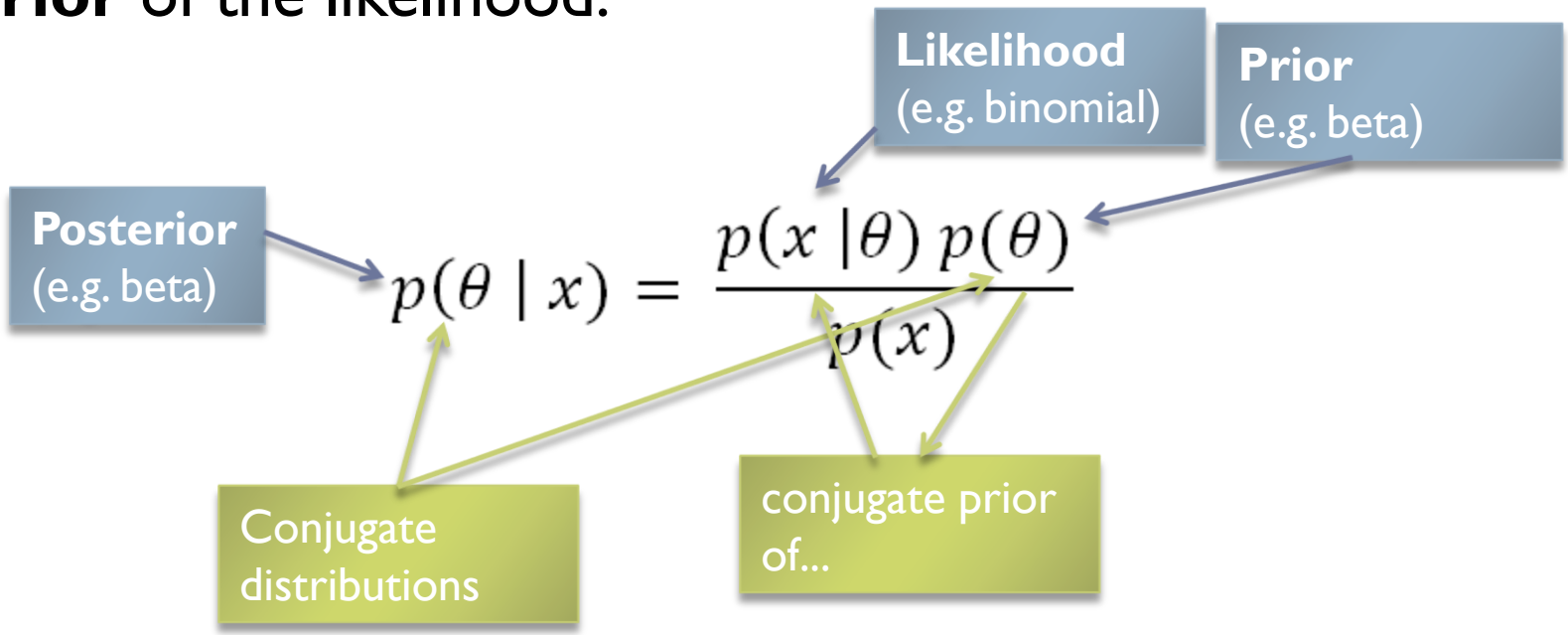
Conjugate Priors

- If we introduce some new evidence (the probability of which is defined by the likelihood), then our updated probability of θ (i.e. the posterior) is based on the combined prior belief and actual observed evidence – where the combined count is obviously just the counts of the two sets of trials.



Conjugate Priors

- So what do we mean by **conjugate prior**?
- If the prior distribution is the same 'kind' as that of the posterior distribution, then these are **conjugate distributions**, and the prior is called the **conjugate prior** of the likelihood.

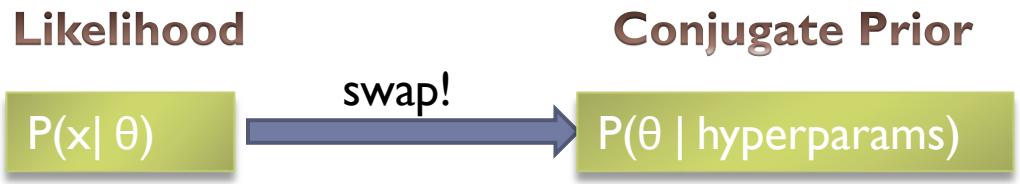


Conjugate Priors

- So what have we done...?
- We have some existing probability model for the likelihood $p(x|\theta)$, but we want to work out the the posterior $p(\theta|x)$.
- This would often involve some difficult integration...
- We choose the **conjugate prior** as the distribution for the prior, with suitable hyperparameters.
- We then know the distribution of the posterior is the same form as the prior (i.e. conjugates), and we can avoid nasty integration.
- The fact the posterior and prior have the same form is useful because we can repeatedly add new evidence (where each time the prior evidence is the evidence we previously saw).

Conjugate Priors

- We previously said that a Dirichlet distribution was the '**reverse**' of the multinomial, and the Beta the '**reverse**' of the Binomial, in the sense that **we just swapped round the parameters and the support**.*
- Because we swapped things round between the two distributions and because we also swapped things round from likelihood $p(x|\theta)$ to prior $p(\theta)$ (except we have some prior evidence rather than actual evidence), we end up multiplying 2 equations of the same form – which allows our simplification tricks (e.g. adding powers) in calculating the posterior.



*There is a MAJOR disclaimer with this. While it is true that we are 'flipping' the support and parameters, we do not necessarily need the same conceptual thing. The Beta distribution is also the conjugate prior of the **Geometric Distribution** for example. While the later is conceptually quite different from the Binomial, the p.m.f. has a similar form, so in this case we can still apply the same tricks.

Conjugate Priors

Likelihood	Conjugate Prior Distribution
Bernoulli	Beta
Binomial	Beta
Multinomial	Dirichlet
Geometric	Beta
Poisson	Gamma
Gaussian	Gaussian

This last one is damn handy in something called a **Kalman Filter**.

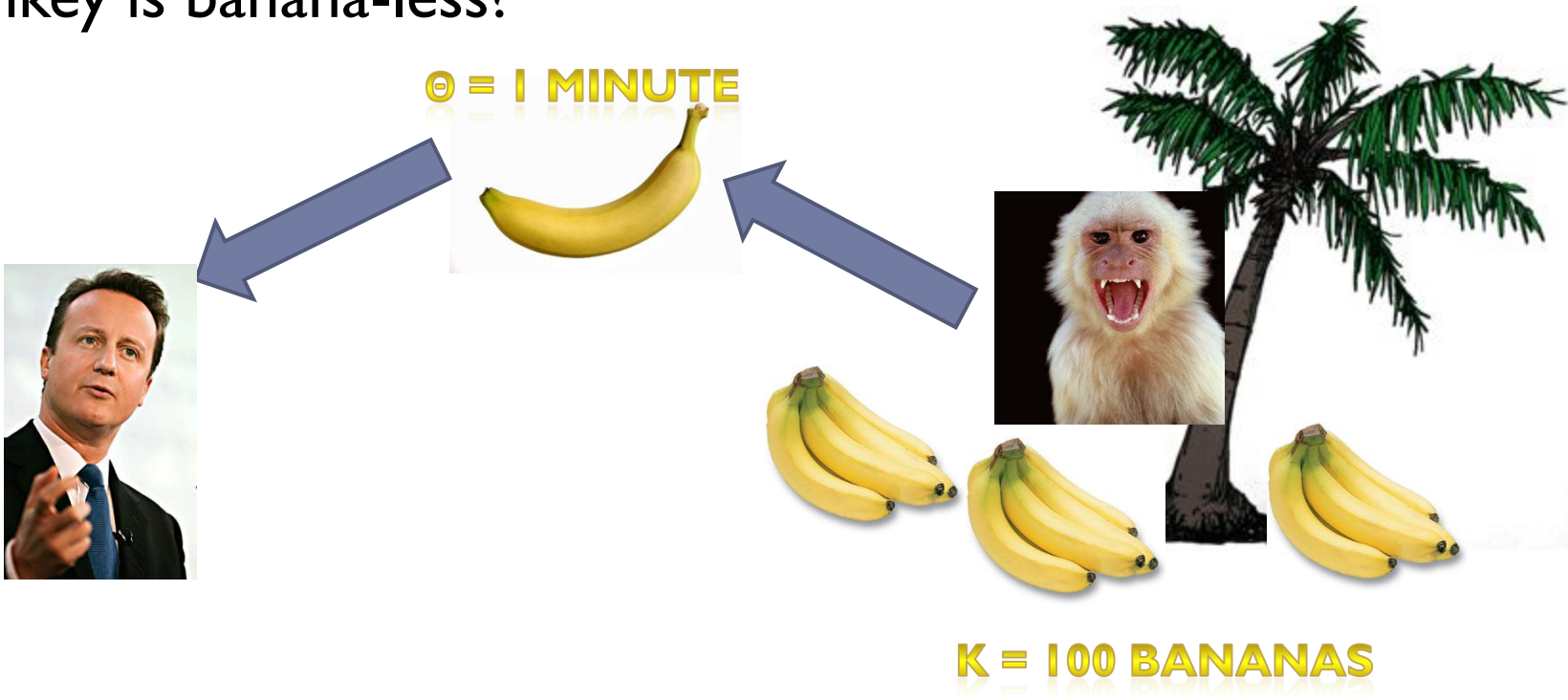
Gamma Distribution

- All that's left is to describe the Gamma Distribution. It can be thought of as the reverse (almost!) of the **Poisson Distribution**.
- Recall that the Poisson Distribution was the number of events in a fixed time given some rate at which events occurs.*
- The Gamma Distribution gives us the total time to perform k events, given the average time of each is θ (i.e. the inverse of the rate).
- So before, we fixed the total time and wanted to know how many events occurred – now we're fixing the number of events and want to know the total time.
- The rate is still a parameter, except with the Gamma Distribution we specify the (average) period θ between events instead.

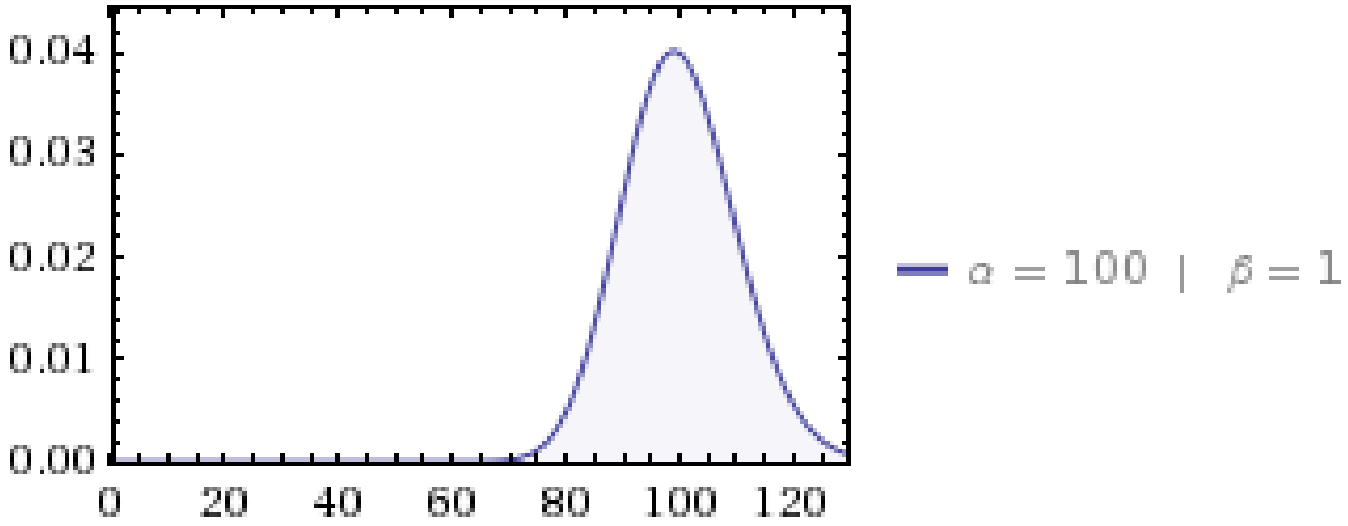
* A **Poisson Process** describes what's going on. We have a fixed average rate/period in which our events occur, each event happens independently of each other, and the Poisson Distribution describes the number of events.

Gamma Distribution

- Example: A rabid monkey has 100 bananas. He lobs a banana at his target on average every 1 minute. What is the distribution that governs the time it takes before the monkey is banana-less?



Gamma Distribution



Rather unremarkably, we expect the average amount of time to be 100 minutes. Obviously:

$$E[X] = k\,\theta$$

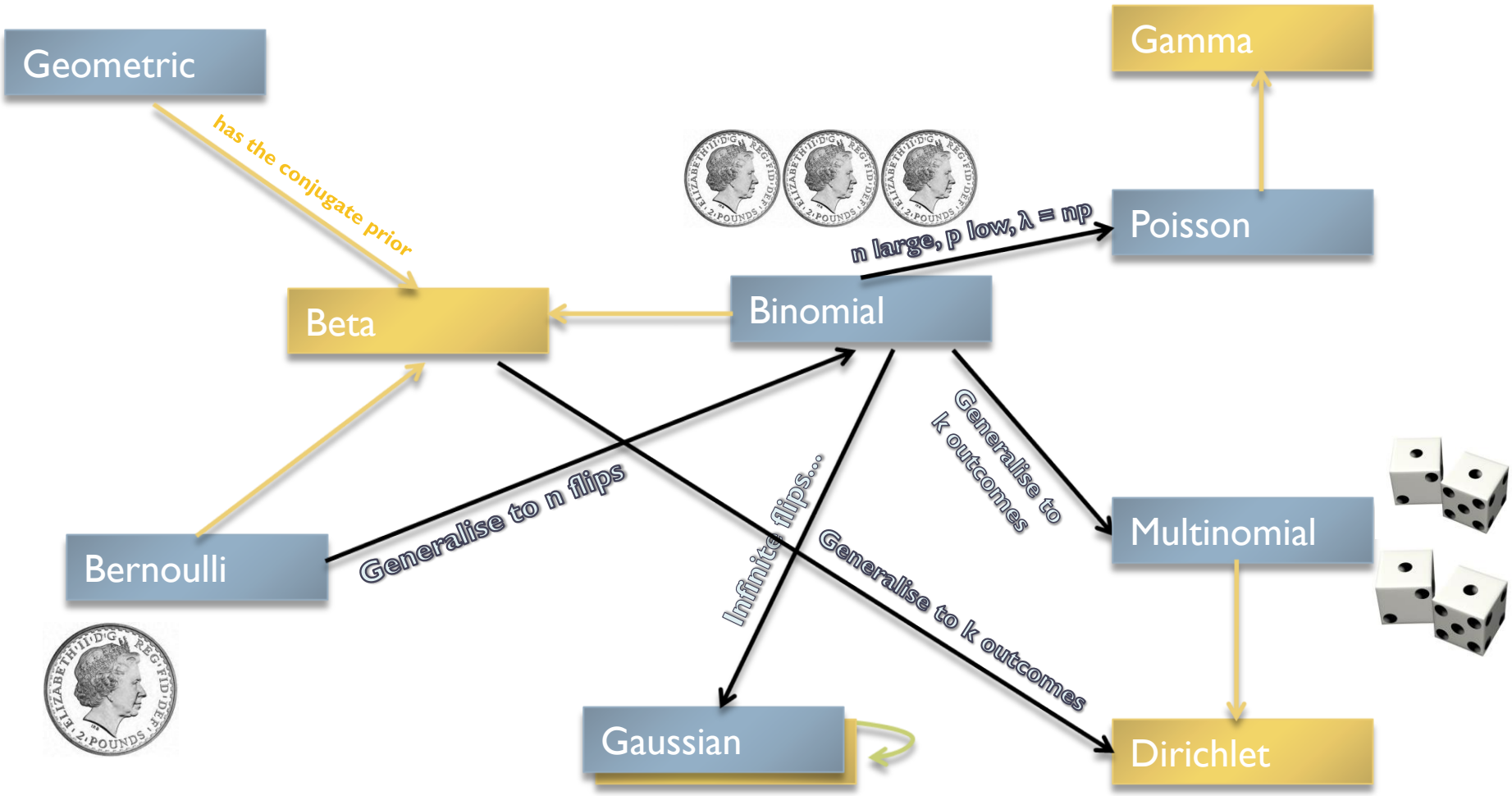
*(But curiously, the **most likely** amount of time is (k-1) ϑ = 99 minutes!)*

The probability density function is:

$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\,\Gamma(k)} \text{ for } x > 0 \text{ and } k,\theta > 0.$$

I just typed in Gamma(100,1) into Wolfram Alpha to get this graph. It will initially think you mean the 'gamma function' rather than the gamma distribution, so click on the 'Use as a probability distribution' link.

Summary (of everything!)



Fini

That's all folks...