

Introduce a running example:

Discrete Case: Single Param Binary vars KNOWN STRUCTURE

Imagine 101 coins in a bucket where each has a different propensity for landing on heads. The first coin has propensity 0.00, the second .01, and the last 1.00. This means that if we toss the second coin many times, the relative frequency with which it landed on heads would approach .01.

Suppose I pick a coin at random -- what probability should I assign to it landing on heads?

If we know what coin we have, this is easy: the relative freq.

If we don't, [let Side be a r.v. with values  $\in \{Heads, Tails\}$ .  
let F be a r.v. whose range consists of the 101 values of the relative frequencies.]

Then  $P(\text{Side} = \text{Heads} | f) = f$ .

where  $f$  denotes  $F=f$ .  
(not used on  $\text{Side} = \text{Heads}$  because the function  $\text{Side} = \text{Heads}$  is implied -  $f$  means any element in space of  $F$ )

If we ~~assign~~ <sup>assign</sup> equal probabilities to all relative frequencies (coins), because we have no reason not to, (principle of indifference) (n items 1/n chance to all), we can represent our probability distribution with this BN:

$P(f) = 1/101 \quad .00 \leq f \leq 1.00$



This is technically an "augmented BN" because it includes a node representing belief about a relative frequency. Such nodes are typically shaded. I'll formally define ABNs later, if needed.

$P(\text{side} = \text{Heads} | f) = f$

So:

$$P(\text{Side} = \text{Heads}) = \sum_{f=0.00}^{1.00} P(\text{Side} = \text{Heads} | f) P(f)$$

$$= \sum_{f=0.00}^{1.00} f \left( \frac{1}{100} \right) = \left( \frac{1}{100 \times 101} \right) \sum_{f=0}^{100} f = \frac{1}{100 \times 101} \times \frac{101 \times 100}{2} = \frac{1}{2}$$

arithmetic series identity  
↓

This is not surprising since the <sup>relative frequencies</sup> ~~propensities~~ are distributed evenly on both sides of 0.5.

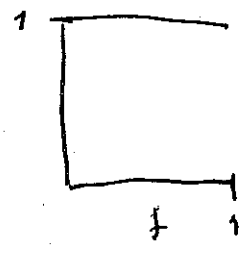
~~0.5~~ is also the r.f. with which heads would occur if we repeatedly sample coins with replacement and toss each coin once. ~~X~~

Done with discrete case: Now, CONTINUOUS (still single param) (still many vars)

Suppose now we have a really big bucket with a continuum of coins, one for every real number  $f$  between 0 and 1. - For each of those real numbers  $f$ , there is a coin with propensity of  $f$  for landing on Heads, and we again pick a coin at random.

Then our probability distribution of the r.v. <sup>whose values are the relative freqs with which we get heads.</sup> is given by the uniform density function,  $p(f) = 1$ .

In this case our probability of landing heads on the first toss is



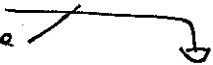
$$P(\text{Side} = \text{Heads}) = \int_0^1 P(\text{side} = \text{Heads} | f) p(f) df$$


$$= \int_0^1 f(1) df = \frac{1}{2}, \text{ which is not surprising.}$$

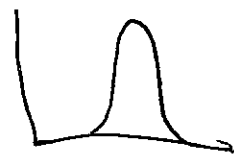
For any problem in which we consider all numbers in  $[0, 1]$  to be equally likely to be the value of the r.f., then we can model our belief with the UDF.

BUT WHAT IF ALL R.F.'s NOT EQUIPROBABLE?

In many cases (if not most), we do not feel all numbers in  $[0,1]$  are equally likely to be the value of a relative frequency.

If I toss a coin from my pocket, I'd think a ~~the~~ r.f. of heads ~~is~~ around .5 to be most probable. So I'd expect a d.f. like this one 

If I asked people if they brush their teeth, I'd expect the most probable r.f. to be around .9, so I'd expect a d.f. like this one: 



The beta density function family provides a natural way of

- 1) quantifying prior beliefs about r.f.'s and
- 2) updating these beliefs in the light of evidence.

# REVIEW: BETA DISTRIBUTION

(4)

First we need Gamma function, which generalizes the factorial function for real and complex numbers.

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \rightarrow \Gamma(x) = (x-1)! \text{ for } x \in \mathbb{N}^*$$

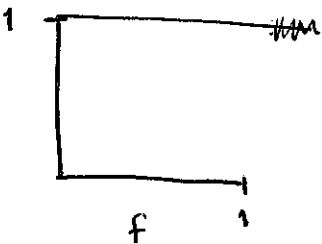
Definition: Beta Density Function

$$\rho(f) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1} (1-f)^{b-1} \quad 0 \leq f \leq 1, \quad a, b \in \mathbb{R}_{>0}$$

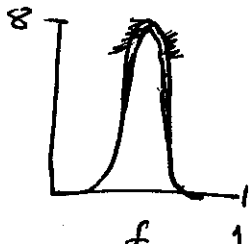
A random variable  $F$  that has this density function is said to have a beta distribution. Refer to beta density function as:

beta( $f; a, b$ ).

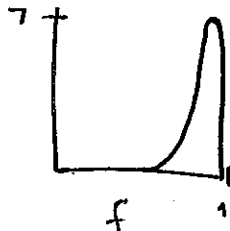
Examples:



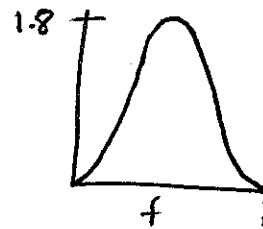
beta( $f; 1, 1$ )



beta( $f; 50, 50$ )



beta( $f; 18, 2$ )



beta( $f; 3, 3$ )

As  $a, b$  increase, more mass centers around  $a/(a+b)$ .

"We see outcome  $a$  happen in  $a+b$  trials."

"Estimation of the relative frequency"

Expectation of random variable  $F$  is

$$E(F) = a/N = a/a+b \quad (\text{proof not shown})$$

Theorem

~~if  $X$  is a random variable with values  $\{1, 2, \dots\}$ , and  $F$  another r.v. s.t.:~~  $P(X=1|F) = F$ , then

$$P(X=1) = E(F) \quad (\text{proof not shown})$$

if we knew for a fact that the r.f. with which  $X=1$  was  $F$ , our belief concerning the occurrence of 1 in the first execution of the experiment would be  $F$ .

Theorem shows that our subjective probability for the first trial is the same as our estimate of relative frequency.

Beta Examples

1) if I toss a coin from my pocket repeatedly, since I feel it highly probable that the r.f. is around .5, I might feel this prior experience is equivalent to having seen 50 heads in 100 trials: beta (f; 50, 50)

$$\text{Thus } P(\text{side} = \text{Heads}) = \frac{50}{50+50} = 0.5$$

Furthermore .5 is our estimate of the r.f. with which we get heads

2) Thumbtack: I feel it should land heads about half the time, but not as certainly as the coin. Say we've "seen" 3/6 encodes our belief: beta (f; 3, 3)

$$\text{Thus } P(\text{side} = \text{Heads}) = \frac{3}{3+3} = 0.5$$

same r.f. estimate.

Different curve.

3) Teeth Brushing encode belief as 18/20 yes: beta (f; 18, 2)

$$P(\text{Teeth} = \text{Brushed}) = \frac{18}{18+2} = 0.9$$

.9 = est. r.f.

How do we learn a relative frequency?

(6) ~~7~~

Recall the coins in the bucket. If we pick 1 at random and toss it, our probability of heads would be 0.5.

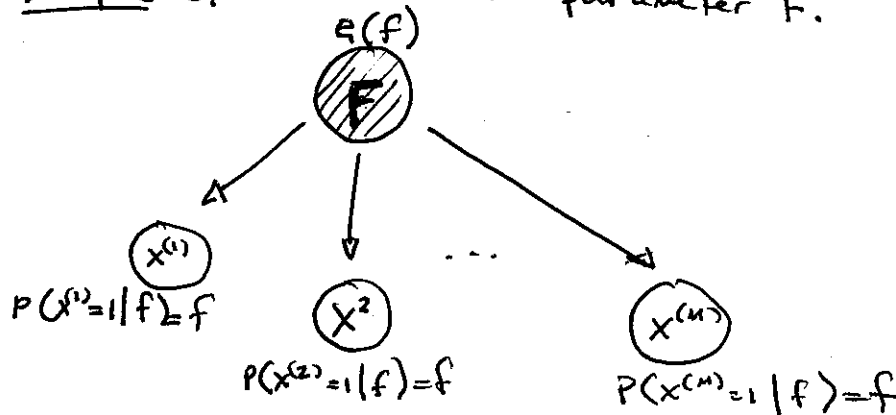
Suppose we've tossed it 20 times and we got 18 heads, 2 tails. Would we continue to assign  $\text{prob} = 0.5$  to the next toss? No -- we'd feel it more probable that the coin is one of those with propensity for heads around 0.9. How to quantify such a change in belief?

### Definition

Suppose  $D = \{X^{(1)}, X^{(2)}, \dots, X^{(M)}\}$  such that each  $X^{(h)}$  has the same space.

Suppose there is a r.v.  $F$  with d.f.  $p$  s.t. all  $X^{(h)}$ 's are mutually independent conditional on  $F$ , and for all  $f \in F$ , all  $X^{(h)}$  have the same probability distribution conditional on  $f$ .

Then  $D$  is called a sample of size  $M$  with parameter  $F$ .



$D$  is a binomial sample if each  $X^{(h)}$  has space  $\{1, 2\}$  and  $F \in [0, 1]$  and  $P(X^{(h)}=1 | f) = f$  for  $1 \leq h \leq M$ .

Theorem:

- 1)  $D$  is a binomial sample of size  $M$  w/ parameter  $F$
- 2) we have a ~~data~~ <sup>data</sup> set of values  $d = \{x^{(1)}, x^{(2)}, \dots, x^{(M)}\}$
- 3)  $s$  is the # of variables in  $d$  equal to 1
- 4)  $t$  is the # of variables in  $d$  equal to 2

Then:  $P(d) = E(F^s [1-F]^t)$

Theorem :

$$p(f|d) = \frac{f^s(1-f)^t p(f)}{E(F^s[1-F]^t)} \quad (\text{proof not shown})$$

Corollary:

if  $p(f) = \text{beta}(f; a, b)$

Then  $p(f|d) = \text{beta}(f; a+s, b+t)$

Given a sample,  $p(f|d)$  is called the updated density function of the parameters relative to the sample and data.

Previous corollary shows that if we update a beta density function relative to a binomial sample, we obtain another beta density function.

Example:

Thumbtack toss 10x 1 → heads 2 → tails,  $x^{(h)}$  → value of  $h^{\text{th}}$  toss. Let density function be  $\text{beta}(f; 3, 3)$  (as before).

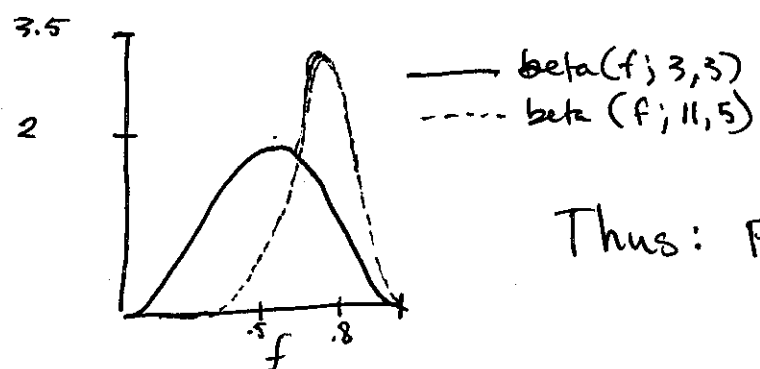
$D = \{x^{(1)}, x^{(2)}, \dots, x^{(10)}\}$  is binomial w/ param  $F$ .

The prior relative to the sample is  $P(x^{(h)}=1) = E(F) = \frac{3}{3+3} = \frac{1}{2}$

Now use  $d = \{1, 1, 2, 1, 1, 1, 1, 1, 2, 1\}$

Then  $a=3, b=3, s=8, t=2$  and the (updated) posterior is

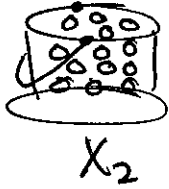
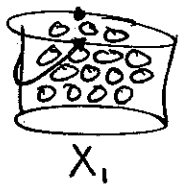
$$p(f|d) = \text{beta}(f; 3+8, 3+2) = \text{beta}(f; 11, 5).$$



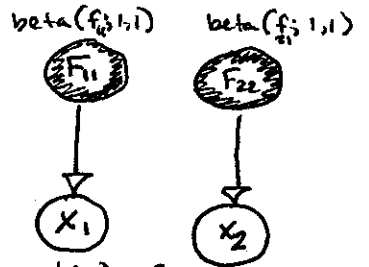
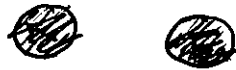
$$\begin{aligned} \text{Thus: } P(x^{(m+1)} = 1 | d) &= E(F|d) \\ &= \frac{a+s}{N+M} \quad (N=a+b) \end{aligned}$$

We've been learning single parameters. What about all? (8)

Recall the bucket of coins, but add another bucket with copies of all the 101 coins. For simplicity, assume uniform distrib. for now.



ABN  
>



$$P(X_1=1 | f_{11}) = f_{11}$$

$$P(X_2=1 | f_{21}) = f_{21}$$

(Augmented BN includes shaded nodes representing beliefs about relative frequencies.)

EX:

Say we sample coins from the buckets, toss them 7x, and get:

Case	$X_1$	$X_2$
1	1	1
2	1	1
3	1	1
4	1	2
5	2	1
6	2	1
7	2	2

Say  $s_{11}$  = # of times  $X_1=1$   
 $t_{11}$  = # of times  $X_1=2$   
 $s_{21}$  = # of times  $X_2=1$   
 $t_{21}$  = # of times  $X_2=2$

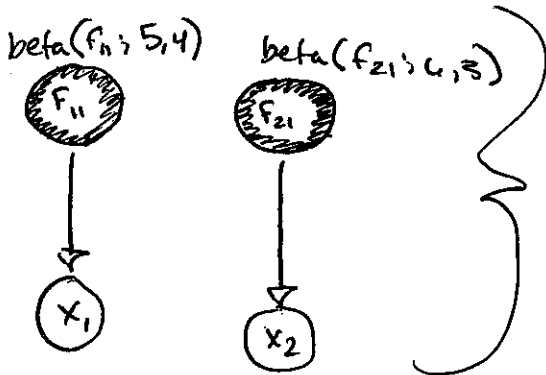
then,

$$p(f_{11} | d) = \text{beta}(f_{11}; a_{11} + s_{11}, b_{11} + t_{11})$$

$$= \text{beta}(f_{11}; 1 + 4, 1 + 3) = \text{beta}(f_{11}; 5, 4)$$

$$p(f_{21} | d) = \text{beta}(f_{21}; a_{21} + s_{21}, b_{21} + t_{21})$$

$$= \text{beta}(f_{21}; 1 + 5, 1 + 2) = \text{beta}(f_{21}; 6, 3)$$



according to this:

$$P(X_1=1, X_2=1) = P(X_1=1)P(X_2=1)$$

$$= \left(\frac{5}{9}\right)\left(\frac{2}{3}\right) = \frac{10}{27}$$

$$P(X_1=1, X_2=2) = \dots = 5/27$$

$$P(X_1=2, X_2=1) = \dots = 8/27$$

$$P(X_1=2, X_2=2) = \dots = 4/27$$

When before they were all  $1/4$



We've been doing only binary case. What about multinomial variables?

Choices:

- 1) quantize data
- 2) use more notation

Dirichlet density function with parameters  $a_1, a_2, \dots, a_r$ ,  $N = \sum_{k=1}^r a_k$   
 $a_1, \dots, a_r \geq 1$

$$p(f_1, f_2, \dots, f_{r-1}) = \frac{\Gamma(N)}{\prod_{k=1}^r \Gamma(a_k)} f_1^{a_1-1} f_2^{a_2-1} \dots f_r^{a_r-1} \quad \begin{matrix} 0 \leq f_k \leq 1 \\ \sum_{k=1}^r f_k = 1 \end{matrix}$$

RV's with  $F_1, F_2, \dots, F_r$  with this density function are said to have the Dirichlet distribution.

( $F_r$  uniquely determined by values of first  $r-1$  variables  $\rightarrow$   $p$  function of  $r-1$  vars)

Dirichlet generalizes Beta.

$$\left. \begin{matrix} E(F_k) = a_k/N \\ P(X=k | f_k) = f_k \end{matrix} \right\} \text{just like beta.}$$

Apply same techniques as we did in Beta.

Continuous Variables?

we use different distributions (i.e. Gaussian)  
 Normal  
 Chi-square

Next Time: STRUCTURE!  
FINALLY!