Meeting 3 and 4 (lecture 2): Bayesian inference - Discrete probability models

Many things about Bayesian inference for discrete probability models are similar to frequentist inference

Discrete probability models:

- Binomial sampling

  Sampling a fix number of trials from a *Bernoulli process*

  A Bernoulli process is a series of trials ($y_1, y_2, \ldots$)
  - where in each trial
    - there two possible outcomes (*success* and *failure*)
    - the probability of success is constant = p
  - where the members of the set of possible sequences $y_{(1)}, \ldots, y_{(M)}$ all with s successes and f failures ($s + f = M$) are ***exchangable***

  The number of successes, $\tilde{r}$ in n trials is binomial distributed

$$P(\tilde{r} = r) = \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{r!(n-r)!} \cdot p^r (1-p)^{n-r} \quad , r = 0,1,\ldots,n$$

- Hypergeometric sampling

  Sampling a fix number n of items (without replacement) from a finite set of N items.

  The finite set of items
  - contains $Np = R$ items of a specific type (“success” item)

  The number of success items, $\tilde{r}$ among the n sampled items is hypergeometric distributed

$$P(\tilde{r} = r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \quad , r = 0,1,\ldots,\min(n,R)$$

- Pascal sampling

Sampling a random number of trials from a Bernoulli process until a predetermined number *r* of successes has been obtained.

The number of trials needed is a random variable $\tilde{n}$ with a Pascal or Negative binomial distribution

$$P\left(\tilde{n} = n \middle| r, p\right) = \binom{n-1}{r-1} p^r \left(1-p\right)^{n-r} \quad , n = r, r+1, \ldots$$

Special case, when *r* = 1: *First success (Fs) distribution*

$$P\left(\tilde{n} = n \middle| p\right) = p\left(1-p\right)^{n-1} \quad , n = 1,2, \ldots$$

Related to the *Geometric distribution*

$$P\left(\tilde{x} = x \middle| p\right) = p\left(1-p\right)^{x} \quad , x = 0,1, \ldots$$

- The Poisson process

A counting process with so-called *independent increments*

The events to be counted appears with an intensity $\lambda(t)$

The number of events appearing in the time interval (t_1, t_2) is Poisson distributed with mean

$$\mu = \int_{t_1}^{t_2} \lambda(t)dt$$

i.e

$$P(\tilde{r} = r | \lambda(t), t_1, t_2) = \frac{\left(\int_{t_1}^{t_2} \lambda(t)dt \right)^r e^{-\int_{t_1}^{t_2} \lambda(t)dt}}{r!} \quad , r = 0,1,\ldots$$

Most common case: $\lambda(t) \equiv \lambda$ (constant) and $t_1 = 0$. $t_2 = t$ (homogeneous process):

$$P(\tilde{r} = r | \lambda, t) = \frac{(\lambda t)^r e^{-\lambda t}}{r!} \quad , r = 0,1,\ldots$$

Bayes' theorem applied to discrete probability distributions

What is observed is what (normally) has a discrete probability distribution.

- In Binomial sampling we observe the number of successes
- In Hypergeometric sampling we observe the number of success items in the sample
- In Pascal sampling we observe how many items need to be sampled
- In a counting experiment we count the number of events in a specified time interval

Hence, the discrete probability distribution applicable rules the *likelihood*.

Bayes' theorem on a very generic form:

$$P(\theta|\text{Data}) \propto L(\theta;\text{Data}) \cdot P(\theta)$$

where P is the probability measure applicable to the parameter θ and $L(\theta;\text{Data})$ is the likelihood of θ in light of the observed Data.

Proportionality constant: $\int_\theta L(\theta;\text{Data}) dP(\theta) = \langle \text{often} \rangle = \int_\theta L(\theta;\text{Data}) \cdot P(\theta) d\theta$

Hence,

for binomial sampling $P(p|n,r) \propto \binom{n}{r} p^r (1-p)^{n-r} \cdot P(p)$

for hypergeometric sampling $P(p|N,n,r) \propto \left[\binom{Np}{r} \binom{N(1-p)}{n-r} \Big/ \binom{N}{n} \right] \cdot P(p)$

for Pascal sampling $P(p|n,r) \propto \binom{n-1}{r-1} p^r (1-p)^{n-r} \cdot P(p)$

for counting in a homogeneous Poisson process $P(\lambda|r,t) \propto \frac{(\lambda t)^r e^{-\lambda t}}{r!} \cdot P(\lambda)$

Extending with *hyper parameters* (ψ)

$$P(\theta|\text{Data},\psi) \propto L(\theta;\text{Data}) \cdot P(\theta|\psi)$$

When θ is continuous and the probability measure is Riemann-Stieltjes integrable (there is a cumulative distribution function)

$$f(\theta|\text{Data},\psi) \propto L(\theta;\text{Data}) \cdot f(\theta|\psi)$$

where *f* stands for a *probability density function* (its form may very well depend on the conditions (ψ and (ψ, Data) respectively)

Exercise 3.24

You feel that $\tilde{p}$, the probability of heads on a toss of a particular coin is either 0.4, 0.5 or 0.6. Your prior probabilities are $P(0.4) = 0.1$, $P(0.5) = 0.7$ and $P(0.6) = 0.2$. You toss the coin three times and obtain heads once and tails twice. What are the posterior probabilities? If you then toss the coin three *more* times and once again obtain heads once and tails twice, what are the posterior probabilities? Also, compute the posterior probabilities by pooling the two samples and revising the original probabilities just once; compare with your previous answers.

Likelihoods from first sample:

$$L\left(p = 0.4; \text{First}\right) = \binom{3}{1} \cdot 0.4^1 \cdot 0.6^2 = 0.432$$

$$L\left(p = 0.5; \text{First}\right) = \binom{3}{1} \cdot 0.5^1 \cdot 0.5^2 = 0.375$$

$$L\left(p = 0.6; \text{First}\right) = \binom{3}{1} \cdot 0.6^1 \cdot 0.4^2 = 0.288$$

Posterior probabilities:

$$P(p|\text{First}) = \frac{L(p;\text{First})\cdot P(p)}{L(0.4;\text{First})\cdot P(0.4)+L(0.5;\text{First})\cdot P(0.5)+L(0.6;\text{First})\cdot P(0.6)}$$

$$\Rightarrow$$

$$P(0.4|\text{First}) = \frac{L(0.4;\text{First})\cdot P(0.4)}{L(0.4;\text{First})\cdot P(0.4)+L(0.5;\text{First})\cdot P(0.5)+L(0.6;\text{First})\cdot P(0.6)}$$

$$= \frac{0.432\cdot 0.1}{0.432\cdot 0.1+0.375\cdot 0.7+0.288\cdot 0.2} = 0.11891$$

$$P(0.5|\text{First}) = \frac{0.375\cdot 0.7}{0.432\cdot 0.1+0.375\cdot 0.7+0.288\cdot 0.2} = 0.7225434$$

$$P(0.6|\text{First}) = \frac{0.288\cdot 0.2}{0.432\cdot 0.1+0.375\cdot 0.7+0.288\cdot 0.2} = 0.1585467$$

Likelihoods from second sample:

$$L(p = 0.4; \text{Second}) = \binom{3}{1} \cdot 0.4^1 \cdot 0.6^2 = 0.432$$

$$L(p = 0.5; \text{Second}) = \binom{3}{1} \cdot 0.5^1 \cdot 0.5^2 = 0.375$$

$$L(p = 0.6; \text{Second}) = \binom{3}{1} \cdot 0.6^1 \cdot 0.4^2 = 0.288$$

These are the same values as with the first sample since the sample outcomes are identical.

Posterior probabilities after second sample;

$$P\left(p\middle|\text{Second},(\text{First})\right)$$

$$=\frac{L\left(p;\text{Second}\right)\cdot P\left(p\middle|\text{First}\right)}{L\left(0.4;\text{Second}\right)\cdot P\left(0.4\middle|\text{First}\right)+L\left(0.5;\text{Second}\right)\cdot P\left(0.5\middle|\text{First}\right)+L\left(0.6;\text{Second}\right)\cdot P\left(0.6\middle|\text{First}\right)}$$

$$P\left(0.4\middle|\text{Second},(\text{First})\right)$$

$$=\frac{L\left(0.4;\text{Second}\right)\cdot P\left(0.4\middle|\text{First}\right)}{L\left(0.4;\text{Second}\right)\cdot P\left(0.4\middle|\text{First}\right)+L\left(0.5;\text{Second}\right)\cdot P\left(0.5\middle|\text{First}\right)+L\left(0.6;\text{Second}\right)\cdot P\left(0.6\middle|\text{First}\right)}$$

$$=\frac{0.432\cdot 0.11891}{0.432\cdot 0.11891+0.375\cdot 0.7225434+0.288\cdot 0.1585467}=0.1395959$$

$$P\left(0.5\middle|\text{Second},(\text{First})\right)$$

$$=\frac{0.375\cdot 0.7225435}{0.432\cdot 0.11891+0.375\cdot 0.7225434+0.288\cdot 0.1585467}=0.7363188$$

$$P\left(0.6\middle|\text{Second},(\text{First})\right)$$

$$=\frac{0.288\cdot 0.1585467}{0.432\cdot 0.11891+0.375\cdot 0.7225434+0.288\cdot 0.1585467}=0.1240853$$

Likelihoods from pooled samples:

$$L(p = 0.4; \text{Pooled}) = \binom{6}{2} \cdot 0.4^2 \cdot 0.6^4 = 0.31104$$

$$L(p = 0.5; \text{Pooled}) = \binom{6}{2} \cdot 0.5^2 \cdot 0.5^4 = 0.234375$$

$$L(p = 0.6; \text{Pooled}) = \binom{6}{2} \cdot 0.6^2 \cdot 0.4^4 = 0.13824$$

Posterior probabilities:

$$P(p|\text{Pooled}) = \frac{L(p;\text{Pooled})\cdot P(p)}{L(0.4;\text{Pooled})\cdot P(0.4)+L(0.5;\text{Pooled})\cdot P(0.5)+L(0.6;\text{Pooled})\cdot P(0.6)}$$

$$\Rightarrow$$

$$P(0.4|\text{Pooled}) = \frac{L(0.4;\text{Pooled})\cdot P(0.4)}{L(0.4;\text{Pooled})\cdot P(0.4)+L(0.5;\text{Pooled})\cdot P(0.5)+L(0.6;\text{Pooled})\cdot P(0.6)}$$

$$= \frac{0.31104\cdot 0.1}{0.31104\cdot 0.1+0.234375\cdot 0.7+0.13824\cdot 0.2} = 0.1395959$$

$$P(0.5|\text{Pooled})$$

$$= \frac{0.234375\cdot 0.7}{0.31104\cdot 0.1+0.234375\cdot 0.7+0.13824\cdot 0.2} = 0.7363188$$

$$P(0.6|\text{Pooled})$$

$$= \frac{0.13824\cdot 0.2}{0.31104\cdot 0.1+0.234375\cdot 0.7+0.13824\cdot 0.2} = 0.1240853$$

Comparison:

$$P(0.4|\text{Second},(\text{First})) = 0.1395959 \qquad P(0.4|\text{Pooled}) = 0.1395959$$

$$P(0.5|\text{Second},(\text{First})) = 0.7363188 \qquad P(0.5|\text{Pooled}) = 0.7363188$$

$$P(0.6|\text{Second},(\text{First})) = 0.1240853 \qquad P(0.6|\text{Pooled}) = 0.1240853$$

Identical results! Expected?

$$P(p|\text{Second},(\text{First}))$$

$$= \frac{L(p;\text{Second})\cdot P(p|\text{First})}{L(0.4;\text{Second})\cdot P(0.4|\text{First}) + L(0.5;\text{Second})\cdot P(0.5|\text{First}) + L(0.6;\text{Second})\cdot P(0.6|\text{First})}$$

$$L(p;\text{Second})\cdot P(p|\text{First})$$

$$= L(p;\text{Second})\cdot \frac{L(p;\text{First})\cdot P(p)}{L(0.4;\text{First})\cdot P(0.4) + L(0.5;\text{First})\cdot P(0.5) + L(0.6;\text{First})\cdot P(0.6)}$$

$$\propto L(p;\text{Second})\cdot L(p;\text{First})\cdot P(p) = \binom{3}{1} p^1 (1-p)^2 \binom{3}{1} p^1 (1-p)^2 \cdot P(p)$$

$$\propto p^2 (1-p)^4 \cdot P(p)$$

$$P(p|\text{Pooled}) = \frac{L(p;\text{Pooled})\cdot P(p)}{L(0.4;\text{Pooled})\cdot P(0.4)+L(0.5;\text{Pooled})\cdot P(0.5)+L(0.6;\text{Pooled})\cdot P(0.6)}$$

$$\propto \binom{6}{2} p^2(1-p)^4 \cdot P(p) \propto p^2(1-p)^4 \cdot P(p)$$

Hence the two ways of computing the posterior probabilities using both samples are always identical

Exercise 3.33

Suppose that you feel that accidents along a particular stretch of highway occur roughly according to a Poisson process and that the intensity of the process is either 2, 3 or 4 accidents per week. Your prior probabilities for these three possible intensities are 0.25, 0.45 and 0.30, respectively. If you observe the highway for a period of three weeks and 10 accidents occur, what are your posterior probabilities?

Likelihoods:

$$L(\lambda = 2; r = 10, t = 3) = \frac{(\lambda \cdot t)^r e^{-\lambda \cdot t}}{r!} = \frac{(2 \cdot 3)^{10} e^{-2 \cdot 3}}{10!} = 0.04130309$$

$$L(\lambda = 3; r = 10, t = 3) = \frac{(3 \cdot 3)^{10} e^{-3 \cdot 3}}{10!} = 0.1185801$$

$$L(\lambda = 4; r = 10, t = 3) = \frac{(4 \cdot 3)^{10} e^{-4 \cdot 3}}{10!} = 0.1048373$$

Posterior probabilities:

$$P\left(\lambda \middle| r=10, t=3\right) = \frac{\frac{(3\lambda)^{10} e^{-3\lambda}}{10!} \cdot P(\lambda)}{\frac{(3\cdot 2)^{10} e^{-3\cdot 2}}{10!} \cdot P(2) + \frac{(3\cdot 3)^{10} e^{-3\cdot 3}}{10!} \cdot P(3) + \frac{(3\cdot 4)^{10} e^{-3\cdot 4}}{10!} \cdot P(4)}$$

$$\left\langle \begin{array}{l} \text{Faculties} \\ \text{and } 3^{10} \text{ terms} \\ \text{cancel out} \end{array} \right\rangle = \frac{\lambda^{10} e^{-3\lambda} \cdot P(\lambda)}{2^{10} e^{-3\cdot 2} \cdot P(2) + 3^{10} e^{-3\cdot 3} \cdot P(3) + 4^{10} e^{-3\cdot 4} \cdot P(4)}$$

$$P\left(\lambda = 2 \middle| r=10, t=3\right) = \frac{2^{10} e^{-3\cdot 2} \cdot 0.25}{2^{10} e^{-3\cdot 2} \cdot 0.25 + 3^{10} e^{-3\cdot 3} \cdot 0.45 + 4^{10} e^{-3\cdot 4} \cdot 0.30} = 0.1085347$$

$$P\left(\lambda = 3 \middle| r=10, t=3\right) = \frac{3^{10} e^{-3\cdot 3} \cdot 0.45}{3^{10} e^{-3\cdot 2} \cdot 0.25 + 3^{10} e^{-3\cdot 3} \cdot 0.45 + 4^{10} e^{-3\cdot 4} \cdot 0.30} = 0.5608804$$

$$P\left(\lambda = 4 \middle| r=10, t=3\right) = \frac{4^{10} e^{-3\cdot 4} \cdot 0.30}{2^{10} e^{-3\cdot 2} \cdot 0.25 + 3^{10} e^{-3\cdot 3} \cdot 0.45 + 4^{10} e^{-3\cdot 4} \cdot 0.30} = 0.3305849$$

Predictive distributions

For an unknown parameter of interest, θ, we would – according to the subjective interpretation of probability

- assign a prior distribution
- upon obtaining data related to θ, compute a posterior distribution

The prior and posterior distributions are used to *make inference* about the unknown θ – *explanatory inference*

We may also be interested in *predictive inference* , i.e. predict data related to θ but not yet obtained

For cross-sectional data the term prediction is mostly used, while for time series data we rather use the term *forecasting*.

Let $y_1, \ldots, y_M, \ldots$ be the set (finite or infinite) of observed values that may be obtained under conditions ruled by the unknown θ.

The uncertainty associated with each observation – i.e. that its value/state cannot be known in advance – is modelled by letting the observed value be the realisation of a random variable $\tilde{y}$ with a probability distribution depending on θ:

$$P\left(\tilde{y} = y_k \middle| \theta\right) = f\left(y_k \middle| \theta\right)$$

Prior-predictive distributions

The prior-predictive distribution of $\tilde{y}$ is the set of marginal probabilities obtained when the dependency on θ is integrated/summed out by weighting the probability mass function $f(y|\theta)$ with the prior distribution of θ:

$$P\left(\tilde{y} = y_k\right) = \begin{cases} \sum_{\theta} f\left(y_k \middle| \theta\right) \cdot P\left(\tilde{\theta} = \theta\right) & \text{if } \theta \text{ assumes a enumerable set of values} \\ \int_{\theta} f\left(y_k \middle| \theta\right) \cdot p(\theta) d\theta & \text{if } \theta \text{ assumes values on a continuous scale} \end{cases}$$

Posterior-predictive distributions

The posterior-predictive distribution of $\tilde{y}$ is the set of marginal probabilities obtained when the dependency on θ is integrated/summed out by weighting the probability mass function $f(y|\theta)$ with the posterior distribution of θ given an already obtained set of observations (Data):

$$P\left(\tilde{y} = y_k \middle| \text{Data}\right)$$

$$= \begin{cases} \sum_{\theta} f\left(y_k|\theta\right) \cdot P\left(\tilde{\theta} = \theta \middle| \text{Data}\right) & \text{if } \theta \text{ assumes a enumerable set of values} \\ \int_{\theta} f\left(y_k|\theta\right) \cdot p\left(\theta \middle| \text{Data}\right) d\theta & \text{if } \theta \text{ assumes values on a continuous scale} \end{cases}$$

Subjective probabilities and the assignments of them

Exercise

- Form groups of 2-3 persons

- Consider the following four events/scenarios

    1. Donald Trump will be re-elected for another four year period in next year's election in USA.
    2. Next winter in Sweden will be colder than normal.
    3. The United Kingdom will leave the European Union on 31 October 2019 ["Brexit"].
    4. The women's world record of 10.49 seconds on 100 metres outdoor (sport of athletics) from 1988 [Florence Griffith-Joyner] will be beaten before next edition of the Olympic Games (2020).

- Each of you should give your personal degree-of-belief in each of these events rounded off to the nearest multiple of 10% and write it down on a piece of paper. Then compare with the rest of the group, discuss the reasons behind your own assignment

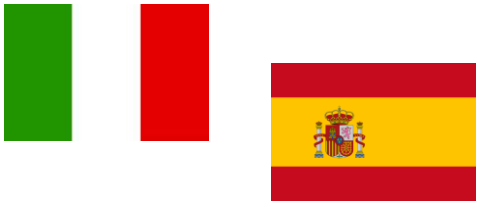
The literature on decision theory/Bayesian analysis usually gives the following method for finding personal probabilities:

- Let E denote the event of which you are supposed to assign your personal probability
- Consider these two lotteries:
  1. You win the amount C with probability p_E
     You win nothing with probability $1 - p_E$
  2. You win the amount C if E happens/is true
     You win nothing if E does not happen/is false
- C is chosen with respect to your economic "status" – lower if you have small resources, larger if you have bigger resources
- The value of p_E that makes you indifferent between these two lotteries is your personal probability of E

Would using this method help you in assigning your personal probabilities of the four events on the previous slide?

Under one and only one set of background information the personal probability of an event must be fix

Assume you would like to assign your personal probability that Italy will beat Spain in a football game. Denote this probability p = Pr (“Italy wins” | I).



Some would say “Well my probability is somewhere between p_1 and p_2” where $p_1 < p_2$ are two numbers between 0 and 1.

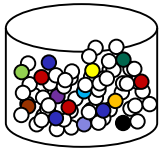
What does such an interval signify?

Is the personal probability a random quantity?

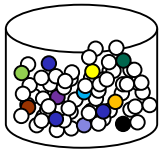
Is p_1 the lowest possible value and p_2 the highest possible value?

Compare with the following scenario:

Assume a pot of 100 balls. You will draw one ball from the pot (only once!) and in front of that assign your probability that the ball drawn will be red.

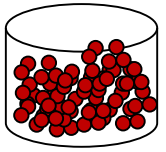


Assume you know that the pot contains no red balls. This constitutes I for your assignment, e.g. denoted by I_0 $\Rightarrow$ Your probability of drawing a red ball should then be 0 (Pr("Red ball" | I_0) = 0 .



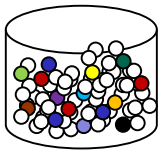
At the same time you know that this probability is *lower* than (or equal to?!) your probability that Italy will beat Spain, i.e. p.

Now, assume you know that all balls in the pot are red, i.e. another I, e.g. denoted by I_{100} . $\Rightarrow$ Your probability of drawing a red ball should now be 1 (Pr("Red ball" | I_{100}) = 1.



At the same time you know that this probability is *higher* than (or equal to?!) your probability that Italy will beat Spain (p).

Now, assume you know that the pot contains x red balls. This constitutes another I for your assignment, e.g. denoted by $I_x \Rightarrow$
Your probability of drawing a red ball should then be $x/100$
$= \Pr(\text{“Red ball”} \mid I_x)$.



If $p = \text{P(”Italy wins”} \mid I)$ is a multiple of 0.01, then there is one and only one particular value of x for which your personal probability for drawing a red ball coincides with p.

You can always reconstruct the pot analogue by extending the number of balls to 1000, 10 000 etc. to fit with the value of p.

If you still would like to use an interval for representing your personal probability?

Does the interval (p_1 , p_2) mean that $\Pr(p_1 \leq p \leq p_2) = 1 - \alpha$ (for α small) ?

…and is ”Pr” still referring to your personal probability measure?

Should there also be intervals for p_1 and p_2 ?

There is a debate on this in the literature, often referring to the issue of a so-called infinite regress (”probability of the probability of the probability …”)

…but compare with ”… of the distribution of hyperparameters of the distribution of hyperparameters of the distribution of parameters.”

When we wish to represent our personal probability as an interval of values, we are actually looking for the *second-order* probability.

When assigning a probability of an event E this is based on the available background information I.

Let us write $I = I(m) = \bigcup_{k=1}^{m} I_k$, where I_1, I_2, … are (mutually exclusive) pieces of background information

Then we would (hopefully) agree on that our assignment of $P(E \mid I(m))$ is a more robust (or at least equally robust) assignment of the probability of E than is $P(E \mid I(n))$ for any $n < m$.

One way of expressing robustness may then be

$$\frac{Pr\left(E \middle| \bigcup_{k=1}^{m} I_k\right)}{Pr\left(E \middle| \bigcup_{k=1}^{\infty} I_k\right)}$$

If this ratio equals 1 there should be no need for an interval representation of the assigned probability of E.

Can we imagine differences between

$$\frac{3}{10}$$

$$\frac{30}{100}$$

$$\frac{3000}{10000}$$

?

Assigning a probability by updating with meagre data

Suppose you are about to assign your personal probability of an event *E*. We may generically denote this probability p_E .

At the outset your background information is $I \Rightarrow p_E = \Pr(E \mid I)$

We can also use odds: $o_E = p_E / (1 - p_E)$

Now, find *a* and *b* such that $p_E = Pr(E|I) = \frac{a}{a+b}$ or $o_E = \frac{a}{b}$

a and *b* then correspond with the parameters of a beta distribution with mean p_E .

If I is meagre, choose a and b as small as possible.

For instance, if your initial assignment is $p_E = 0.15$ based on meagre I ,

- use the fact that $0.15 = 15/100 = 15/(85+15)$
- find the greatest common divisor of 15 and 85 $\Rightarrow 5 \Rightarrow 0.15 = 3/20$
- choose $a = 3$ and $b = 17$

If I is substantial, find a multiplier for a and b that correspond with the extension of I.

For instance, if your initial assignment is $p_E = 0.15$,

- $a = 2 \times 3 = 6$, $b = 2 \times 17 = 24 \Rightarrow 6/40$
- $a = 10 \times 3 = 30$, $b = 10 \times 17 = 170 \Rightarrow 30/200$

Now, assume you extend your background information with some data providing a relative frequency for E : $f_E = n_E / n$

Since the likelihood $L(p)$ of p given your data, is proportional to

$$p^{n_E} \cdot (1-p)^{n-n_E}$$

the beta distribution is the conjugate family of prior/posterior distributions (we'll return to this concept later)

Hence, the posterior distribution from updating with data is beta with parameters $a' = a + n_E$ and $b' = b + n - n_E$

… and the updated assignment of p_E becomes

$$p_E = Pr(E|I,n,E) = \frac{a'}{a'+b'} = \frac{a+n_E}{a+n_E+b+n-n_E} = \frac{a+n_E}{a+b+n}$$

The balance between a meagre or substantial I and meagre or substantial data is built-in.