Meeting 2 (lecture 2): Probability and likelihood, Bayesian inference, subjective probabilities

Purpose: To continue the repetition and extension of probability calculus.

Example:

Assume a method for detecting a certain kind of dye on banknotes is such that

• it gives a positive result (detection) in 99 % of the cases when the dye is present, i.e. the proportion of false negatives is 1%

• it gives a negative result in 98 % of the cases when the dye is absent, i.e. the proportion of false positives is 2%

The presence of dye is rare: prevalence is about 0.1 %

Assume the method has given positive result for a particular banknote.

What is the conditional probability that the dye is present?

Solution:

Let A = "Dye is present" and B = "Method gives positive result"

What about I ?

• We must assume that the particular banknote is as equally likely to be exposed to dye detection as any banknote in the population of banknotes.

• **Is that a realistic assumption?**

Now, $P(A) = 0.001; \quad P(B|A) = 0.99; \quad P\left(B\left|\overline{A}\right.\right) = 0.02$

Applying Bayes' theorem gives

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P\left(B\left|\overline{A}\right.\right) \cdot P(\overline{A})} =$$

$$= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} =$$

Odds and Bayes' theorem on odds form

The *odds* for an event A "is" a quantity equal to the probability:

$$Odds(A) = \frac{P(A)}{P(\overline{A})} = \frac{P(A)}{1 - P(A)} \quad \Rightarrow \quad P(A) = \frac{Odds(A)}{Odds(A) + 1}$$

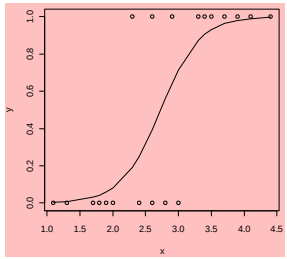
Why two quantities for the same thing?

Example: An "epidemiological" model

Assume we are trying to model the probability p of an event (i.e. the prevalence of some disease).

The *logit link* between p and a set of k explanatory variables $x_1, x_2, \ldots, x_k$ is

$$\text{logit}(p) = \ln \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$



This link function is common in *logistic regression analysis*.

Note that we are modelling the natural logarithm of the odds instead of modelling p.

As the odds can take any value between 0 and ∞ the logarithm of the odds can take any value between $-\infty$ and ∞ ➔ Makes the model practical.

Conditional odds $$Odds(A|B) = \frac{P(A|B)}{P(\overline{A}|B)}$$

express the *updated* belief that *A* holds when we take into account that *B* holds

Like probabilities, all odds are conditional if we include background knowledge *I* as our basis for the calculations. $$Odds(A|I) = \frac{P(A|I)}{P(\overline{A}|I)}; \quad Odds(A|B,I) = \frac{P(A|B,I)}{P(\overline{A}|B,I)}$$

The odds ratio: $$OR = \frac{Odds(A|B,I)}{Odds(A|I)} = \frac{\frac{P(A|B,I)}{P(\overline{A}|B,I)}}{\frac{P(A|I)}{P(\overline{A}|I)}}$$

expresses *how* the belief that *A* holds updates when we take into account that *B* holds.

Now $$\underline{Odds(A|B,I)} = \frac{P(A|B,I)}{P(\overline{A}|B,I)} = \frac{\frac{P(B|A,I)\cdot P(A|I)}{P(B|I)}}{\frac{P(B|\overline{A},I)\cdot P(\overline{A}|I)}{P(B|I)}} =$$
$$= \frac{P(B|A,I)}{P(B|\overline{A},I)} \cdot \frac{P(A|I)}{P(\overline{A}|I)} = \underline{\frac{P(B|A,I)}{P(B|\overline{A},I)} \cdot Odds(A|I)}$$

"*Bayes' theorem on odds form*"

The ratio $\frac{P(B|A,I)}{P(B|\overline{A},I)}$

is a special case of what is called a *likelihood ratio* (the concept of "likelihood" will follow)

$$LR = \frac{P(B|A,I)}{P(B|C,I)}$$

where we have substituted C for $\bar{A}$ and we no longer require A and C to be complementary events (not even mutually exclusive).

$$\frac{P(A|B,I)}{P(C|B,I)} = \frac{P(B|A,I)}{P(B|C,I)} \cdot \frac{P(A|I)}{P(C|I)}$$

always holds, but the ratios involved are not always odds

"The updating of probability ratios when a new event is observed goes through the likelihood ratio based on that event."

Probability and Likelihood – Synonyms?

An event can be *likely* or *probable*, which for most people would be the same. Yet, the definitions of probability and likelihood are different.

In a simplified form:

- The probability of an event measures the degree of belief that this event is true and is used for reasoning about not yet observed events
- The likelihood of an event is a measure of how likely that event is in light of another *observed* event
- Both are objected to probability calculus

More formally…

Consider the *unobserved* event *A* and the *observed* event *B*.

There are probabilities for both representing the degrees of belief for these events in general: $P(A|I)$, $P(B|I)$

However, as *B* is observed we might be interested in

$P(A|B,I)$

which measures the *updated* degree of belief that *A* is true once we know that *B* holds. Still a probability, though.

How interesting is

$P(B|A,I)$

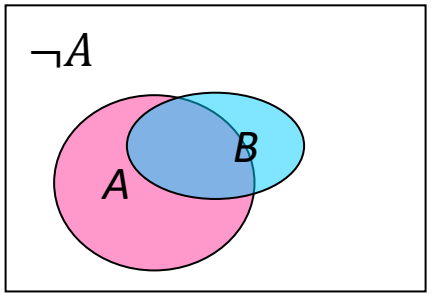
?

$P(B \mid A, I)$ might look meaningless to consider as we have actually observed B.

However, it says something about A.

We have observed B and if A is relevant for B we may compare $P(B \mid A, I)$ with $P(B \mid \bar{A}, I)$.

Now, even if we have not observed A or $\bar{A}$, one of them must be true (as a consequence of A and B being relevant for each other).



If $P(B \mid A, I) > P(B \mid \bar{A}, I)$ we may conclude that A is more *likely* to have occurred than is $\bar{A}$, or better phrased:

"A is a better *explanation* for why B has occurred than is $\bar{A}$".

$P(B \mid A, I)$ is called the *likelihood* of A given the observed B (and $P(B \mid \bar{A}, I)$ is the likelihood of $\bar{A}$).

Note! This is different from the conditional probability of A given B: $P(A \mid B, I)$.

Potential danger in mixing things up:

When we say that an event is the more likely one in light of data we do not say that this event has the highest probability.

Using the likelihood as a measure of how likely is an event is a matter of *inference to the best explanation*.

Logics: Implication:

$$A \rightarrow B$$

- If A is true then B is true, i.e. $P(B \mid A, I) \equiv 1$
- If B is false then A is false, i.e. $P(A \mid \overline{B}, I) \equiv 0$
- If B is true we cannot say anything about whether A is true or not (implication is different from equivalence)

"Probabilistic implication":

$$A \xrightarrow{P} B$$

- If A is true then B *may* be true, i.e. $P(B|A,I) > 0$
- If B is false the A may still be true, i.e. $P(A|\overline{B},I) > 0$
- If B is true then we may decide which of A and $\bar{A}$ is the best explanation

Inference to the best explanation:

- B is observed
- $A_1, A_2, \ldots, A_m$ are potential alternative explanations to B
- If for each $j \neq k$ $P(B \mid A_k, I) > P(B \mid A_j, I)$ then A_k is considered the best explanation for B and is provisionally accepted

Bayes' theorem – different forms

The original "insight" by Thomas Bayes: $$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

"on ordinal form", probabilities of sets, simple version: $$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

"on ordinal form", probabilities of sets, complete version: $$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_j P\left(B\middle|A_j\right) \cdot P\left(A_j\right)}$$

"on odds form", probabilities of sets: $$\frac{P(A|B)}{P(\bar{A}|B)} = \frac{P(B|A)}{P(B|\bar{A})} \cdot \frac{P(A)}{P(\bar{A})}$$

"on ordinal form", for probability density functions: $$f''(y|x) = \frac{f(x|y) \cdot f'(y)}{\int f(x|z) \cdot f'(z) dz}$$

"on odds form", likelihoods from continuous data: $$\frac{P(A|\boldsymbol{x})}{P(\bar{A}|\boldsymbol{x})} = \frac{\int_{"\boldsymbol{y} \in A"} f(\boldsymbol{x}|\boldsymbol{y}) p(\boldsymbol{y}|A)}{\int_{"\boldsymbol{y} \in \neg A"} f(\boldsymbol{x}|\boldsymbol{y}) p(\boldsymbol{y}|\bar{A})} \cdot \frac{P(A)}{P(\bar{A})}$$

The generic form

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

where P is the probability measure applicable to the parameter (or variable) $\boldsymbol{\theta}$, $L(\boldsymbol{\theta};\text{Data})$ is the likelihood of $\boldsymbol{\theta}$ in light of the observed Data, and $\boldsymbol{\psi}$ represents potential hyperparameters.

Proportionality constant: $\int_{\boldsymbol{\vartheta}} L(\boldsymbol{\vartheta};\text{Data}) dP(\boldsymbol{\vartheta})$

When $\boldsymbol{\theta}$ is continuous-valued and the probability measure is Riemann-Stieltjes integrable (there is a cumulative distribution function)

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

where f stands for a *probability density function* (its form may very well depend on the conditions ($\boldsymbol{\psi}$ and ($\boldsymbol{\psi}$, Data) respectively)

Applications to different sampling models

A Bernoulli process is a series of trials ($y_1, y_2, \ldots$)

- where in each trial
  - there are 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***

- Binomial sampling:

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

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

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

  Bayes' theorem for making inference on p:

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

  Common to assume $P(p)$ to follow a beta distribution

- 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

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

  Bayes' theorem for making inference on p (or on R):

$$P(p|N, n, r) \propto \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \cdot P(p)$$

- 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

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

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

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

  Related to the *Geometric distribution*

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

  Bayes' theorem for making inference on *p*:

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

Application to the Poisson process

A counting process with so-called *independent increments*

The events to be counted ($\tilde{r}$) appears with an intensity $\lambda(t)$

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

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

i.e

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

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

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

Bayes' theorem for making inference on λ:

$$P(\lambda|r, t) \propto \frac{(\lambda \cdot t)^r \cdot e^{-\lambda \cdot t}}{r!} \cdot P(\lambda)$$

Exercise

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:

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

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

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

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

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

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:

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

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 or density function $f(y|\theta)$ with the prior distribution of θ:

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

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 or density function $f(y|\theta)$ with the posterior distribution of θ given an already obtained set of observations (Data):

$$P(\tilde{y} = y_k) = \begin{cases} \sum_{\theta} f(y_k|\theta) \cdot P(\tilde{\theta} = \theta|\text{Data}) & \theta \text{ assumes an enumerable set of values} \\ \int_{\theta} f(y_k|\theta) \cdot f''(\theta|\text{Data})\, d\theta & \theta \text{ assumes values on a continuous scale} \end{cases}$$

Subjective probabilities and the assignments of them

Example

- Consider the following four events/scenarios

    1. Kamala Harris will win the election for president in the US 2024 .
    2. The number of bears shot in Sweden so far this year is more than 400.
    3. There will be extensive actions to globally vaccinate people against monkeypox during 2025.
    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 (2028).

- Try to 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.

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
- The value of p_E that makes you indifferent between these two lotteries is your personal probability of E

In practice, you start with $p_E = 0.5$. If lottery 1 is preferred to lottery 2, your $P(E)$ is less than 0.5. If lottery 2 is preferred to lottery 1, then your $P(E)$ is greater than 0.5. Then, continue with $p_E = 0.25$ or $p_E = 0.75$ depending on which lottery was preferred with $p_E = 0.5$, etc.

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 $p = P(\text{“Italy wins”} \mid I)$.

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.

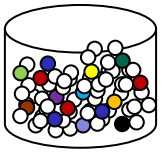
What does such an interval signify?

Is the personal probability a random quantity?

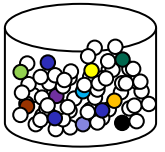
Is p_1 the lowest possible value and p_2 the highest possible value?

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.

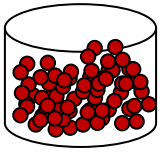


Assume you know that the pot contains no red balls. This constitutes I for your assignment, e.g. denoted by I_0 $\Rightarrow$ Your probability of drawing a red ball should then be 0. [$P(\text{“Red ball”} \mid I_0) = 0$]



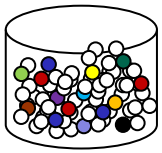
At the same time you know that this probability is *lower* than (or equal to?!) your probability that Italy will beat Spain, i.e. p.

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. [$P(\text{“Red ball”} \mid I_{100}) = 1$]



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 $x/100$
$= P(\text{“Red ball”} \mid I_x)$.



If $p = P(\text{”Italy wins”} \mid I)$ 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 resolution 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 $P(p_1 \leq p \leq p_2) = 1 - \alpha$ (for α small) ?

…and is *"P"* 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 $I = I(n) = \bigcup_{k=1}^{n} I_k$, where I_1, I_2, … are (mutually exclusive) pieces of background information

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

One way of expressing robustness may then be

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

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 $I \Rightarrow p_E = \mathrm{P}(E \mid I)$

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

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

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 corresponds with the extension of I.

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

- $a = 2 \times 3 = 6, \quad b = 2 \times 17 = 34 \quad \Rightarrow \quad 6/40$
- $a = 10 \times 3 = 30, \quad b = 10 \times 17 = 170 \quad \Rightarrow \quad 30/200$

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

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 (using the posterior mean) becomes

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

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