Meeting 2 (lecture 2): Bayesian inference, subjective probabilities

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: $$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)}$$

"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 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:

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

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. US president Joe Biden will remain in office for his entire term.
    2. The number of bears shot in Sweden so far this year is more than 400.
    3. New restrictions due to the Covid-19 virus spread will be imposed in Sweden before Christmas 2022.
    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 (2024).

- 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.

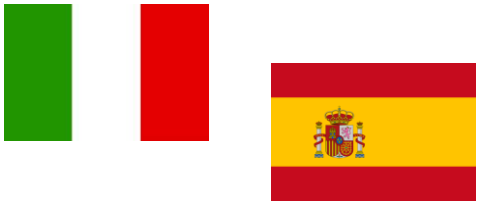
The literature on decision theory/Bayesian analysis usually gives the following method for finding personal probabilities:

- 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
- 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 $p = P(\text{“Italy wins”} \mid I)$.



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.

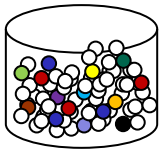
What does such an interval signify?

Is the personal probability a random quantity?

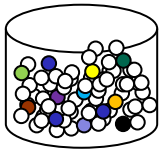
Is p_1 the lowest possible value and p_2 the highest possible value?

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.

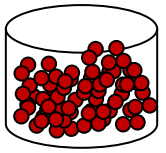


Assume you know that the pot contains no red balls. This constitutes I for your assignment, e.g. denoted by I_0 $\Rightarrow$ Your probability of drawing a red ball should then be 0. [$P(\text{“Red ball”} \mid I_0) = 0$]



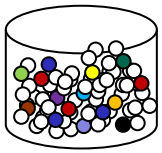
At the same time you know that this probability is *lower* than (or equal to?!) your probability that Italy will beat Spain, i.e. p.

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 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 $P(p_1 \leq p \leq p_2) = 1 - \alpha$ (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

$$\frac{P(E \mid \bigcup_{k=1}^{n} I_k)}{P(E \mid \bigcup_{k=1}^{\infty} I_k)}$$

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 $p_E = P(E|I) = \frac{a}{a+b}$ 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 corresponds with the extension of I.

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

- $a = 2 \times 3 = 6$, $b = 2 \times 17 = 34$ $\Rightarrow$ 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

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.