Meeting 4:
Exponential class of distributions,
Interpretation of priors
Other theories for modelling belief

Notation used for probability mass functions, probability density functions, and random variables

In the current course book by Peterson, probability mass functions (pmf), probability density functions (pdf) and random variables are not explicitly taken up.

But we need them in this course ☺.

To simplify reusing course material from previous years, we therefore adopt the notation form the previous course book (by Winkler).

Random variables: Put tilde (~) above the observable or non-observable quantity
e.g. $\tilde{x}, \tilde{\theta}$

Prior pmf, pdf:	$f'(\cdot)$ [e.g. $f'(\theta)$]
Posterior pmf, pdf:	$f''(\cdot \,\vert\text{"Data"})$ [e.g. $f''(\theta\vert\boldsymbol{x})$]
pmf/pdf of observations:	$f(x\vert\theta)$
Likelihood:	$f(\text{"Data"}\vert\,\cdot)$ [e.g. $f(\boldsymbol{x}\vert\theta)$]
Marginal likelihood, predictive function:	$f(\boldsymbol{x})$, $f(x_{n+1}\vert\boldsymbol{x})$, …

The exponential class of distributions

A (family) of probability distribution(s) belong(s) to the *k*-parameter exponential class of distributions if the probability density (or mass) function can be written:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(\boldsymbol{x})+C(\boldsymbol{x})+D(\boldsymbol{\theta})}$$

where

- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- $A_1(\boldsymbol{\theta}), \ldots, A_k(\boldsymbol{\theta})$ and $D(\boldsymbol{\theta})$ are functions of the parameter $\boldsymbol{\theta}$ only (and not of $\boldsymbol{x}$)
- $B_1(\boldsymbol{x}), \ldots, B_k(\boldsymbol{x})$ and $C(\boldsymbol{x})$ are functions of $\boldsymbol{x}$ only (and not of $\boldsymbol{\theta}$)

Boldface indicates that observations and/or parameters can be multidimensional.

Canonical form: $A_j(\boldsymbol{\theta}) = \theta_j$

Examples

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(\boldsymbol{x})+C(\boldsymbol{x})+D(\boldsymbol{\theta})}$$

Two parameter Gamma distribution (univariate), shape and rate parameterization:

$$f(x|\theta) = f(x|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad ; \; x \geq 0$$

$$= \cdots = e^{\alpha \ln x + \beta(-x) - \ln x + \alpha \ln \beta - \ln \Gamma(\alpha)}$$

$A_1(\alpha,\beta) = \alpha$

$A_2(\alpha,\beta) = \beta$

$B_1(x) = \ln x$

$B_2(x) = -x$

$C(x) = -\ln x$

$D(\alpha,\beta) = \alpha \ln \beta - \ln \Gamma(\alpha)$

Poisson distribution:

$$f(x|\theta) = f(x|\mu) = \frac{\mu^x}{x!} e^{-\mu} = e^{(\ln \mu)\cdot x - \ln x! - \mu} = e^{(\ln \mu)\cdot x - \ln \Gamma(x+1) - \mu} \; ; \; x = 0,1,\ldots$$

$A(\mu) = \ln \mu$

$C(x) = -\ln \Gamma(x+1)$

$B(x) = x$

$D(\mu) = -\mu$

Conjugate families of distributions when the likelihood belongs to the exponential class

pdf (or pmf) of sample point distribution : $f(\boldsymbol{x}|\boldsymbol{\theta}) = e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(\boldsymbol{x})+C(\boldsymbol{x})+D(\boldsymbol{\theta})}$

Likelihood from sample of *n* observations:

$$\prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(\boldsymbol{x}_i)+C(\boldsymbol{x}_i)+D(\boldsymbol{\theta})}$$

$$= e^{\sum_{i=1}^{n}\left(\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(\boldsymbol{x}_i)+C(\boldsymbol{x}_i)+D(\boldsymbol{\theta})\right)}$$

$$= e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\underbrace{\sum_{i=1}^{n} B_j(\boldsymbol{x}_i)}+\underbrace{\sum_{i=1}^{n} C(\boldsymbol{x}_i)}+n\cdot D(\boldsymbol{\theta})}$$

$$B_j'(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) \qquad C'(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})$$

Hence the multivariate array $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ with independent marginal distributions all with density $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ also belongs to the exponential class.

Now, mimic the structure of the exponential class (for the marginal distributions or the likelihood) and define the prior density for $\boldsymbol{\theta}$ as

$$f'\left(\boldsymbol{\theta}\middle|\alpha_1,\ldots,\alpha_k,\alpha_{k+1}\right)$$

$$= e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\cdot\alpha_j+\alpha_{k+1}\cdot D(\boldsymbol{\theta})+K(\alpha_1,\ldots,\alpha_k,\alpha_{k+1})}$$

$$\propto e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\cdot\alpha_j+\alpha_{k+1}\cdot D(\boldsymbol{\theta})}$$

where $\alpha_1, \ldots, \alpha_{k+1}$ are the hyperparameters of this prior distribution and $K(\,)$ is a function of $\alpha_1, \ldots, \alpha_{k+1}$ only .

Then the posterior becomes

$$f''\left(\boldsymbol{\theta}\middle|\{\boldsymbol{x}\},\alpha_1,\dots,\alpha_k,\alpha_{k+1}\right)=f''\left(\boldsymbol{\theta}\middle|\boldsymbol{x}_1,\dots,\boldsymbol{x}_n;\alpha_1,\dots,\alpha_k,\alpha_{k+1}\right)$$

$$\propto \underbrace{\prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta})}_{\text{likelihood}} \cdot f'\left(\boldsymbol{\theta}\middle|\alpha_1,\dots,\alpha_k,\alpha_{k+1}\right)$$

$$= e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta}) \sum_{i=1}^{n} B_j(\boldsymbol{x}_i)+\sum_{i=1}^{n} C(\boldsymbol{x}_i)+n\cdot D(\boldsymbol{\theta})} \cdot e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\cdot\alpha_j+\alpha_{k+1}\cdot D(\boldsymbol{\theta})+K(\alpha_1,\dots,\alpha_k,\alpha_{k+1})}$$

$$= e^{\sum_{i=1}^{n} C(\boldsymbol{x}_i)} e^{K(\alpha_1,\dots,\alpha_k,\alpha_{k+1})} e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n} B_j(\boldsymbol{x}_i)+\alpha_j\right)+(n+\alpha_{k+1})\cdot D(\boldsymbol{\theta})}$$

$$\propto e^{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n} B_j(\boldsymbol{x}_i)+\alpha_j\right)+(n+\alpha_{k+1})\cdot D(\boldsymbol{\theta})}$$

i.e. the posterior distribution is of the same form as the prior distribution but with hyperparameters

$$\alpha_1+\sum_{i=1}^{n} B_1(\boldsymbol{x}_i),\dots,\alpha_k+\sum_{i=1}^{n} B_k(\boldsymbol{x}_i),\alpha_{k+1}+n$$

instead of

$$\alpha_1,\dots,\alpha_k,\alpha_{k+1}$$

Example

Data is Poisson distributed $\Rightarrow \quad f(x|\mu) = \frac{\mu^x}{x!} e^{-\mu} = e^{(\ln \mu)\cdot x - \ln x! - \mu}$

Mimic structure to obtain the prior density for θ:

$$f'(\mu) = e^{ln(\mu)\cdot\alpha_1 + \alpha_2\cdot(-\mu) + K(\alpha_1,\alpha_2)} = \mu^{\alpha_1} \cdot e^{-\alpha_2\cdot\mu} \cdot C(\alpha_1, \alpha_2) \propto \mu^{\alpha_1} \cdot e^{-\alpha_2\cdot\mu}$$

Hence, the prior must be a two-parameter Gamma distribution.

Some common families (within or outside the exponential family):

Conjugate prior	*Sample distribution*	*Posterior*
Beta $\pi \sim Beta(\alpha, \beta)$	Binomial $X \sim Bin(n, \pi)$	Beta $\pi \mid x \sim Beta(\alpha + x, \beta + n - x)$
Normal $\mu \sim N(\varphi, \tau^2)$	Normal, known σ^2 $X_i \sim N(\mu, \sigma^2)$	Normal $\mu \mid \bar{x} \sim N\left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\varphi + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{x}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$
Gamma $\lambda \sim Gamma(\alpha, \beta)$	Poisson $X_i \sim Po(\lambda)$	Gamma $\lambda \mid \sum x_i \sim Gamma(\alpha + \sum x_i, \beta + n)$
Pareto $p(\theta) \propto \theta^{-\alpha};\ \theta \geq \beta$	Uniform $X_i \sim U(0, \theta)$	Pareto $q(\theta; \mathrm{x}) \propto \theta^{-(\alpha+n)};\ \theta \geq \max\left(\beta, x_{(n)}\right)$

Interpretation of prior distributions

Prior distribution for a proportion **(*mean of a beta distribution,*** see lecture notes for Meeting 2!**)**

Prior distribution for the mean *of a population with continuous variation*

Very often we have reasons to work with normally distributed data to make inference about the population mean $\tilde{\mu}$.

If the population variance is (assumed to be) known = σ^2, we can use the normal distribution as a conjugate prior distribution.

From sampling theory we know that – setting aside finite population corrections – the variance of the sample mean is the population variance divided by the sample size

$$Var(\bar{\tilde{y}}|\sigma^2, n) = \frac{\sigma^2}{n}$$

If σ'^2 represents the prior variance of the unknown $\tilde{\mu}$ define a new parameter n' as

$$n' = \frac{\sigma^2}{\sigma'^2}$$

Hence, $$\sigma'^2 = \frac{\sigma^2}{n'}$$

This can be interpreted as the variance σ'^2 of a sample mean based on n' observations taken from the population with population variance σ^2.

n' then plays the role of the size of a virtual sample taken from the population on which the prior knowledge stems.

Note that it is not necessary for n' to be integer-valued, even if it often suffices to approximate with an integer.

For the prior and posterior distribution we may thus write

$$\tilde{\mu} \sim N\left(m', \frac{\sigma^2}{n'}\right) \qquad \tilde{\mu}|\boldsymbol{y} \sim N\left(m'', \frac{\sigma^2}{n''}\right) \qquad \text{where } n'' = \frac{\sigma^2}{\sigma''^2} = n' + n$$

Quick looks at other theories for understanding beliefs

Consider the following case (from forensic science):

An attempt of burglary is recorded on a CCTV camera and it stands clear that the perpetrator is using a crowbar when trying to break the door to the premises (target of the intended burglary). The face of the perpetrator cannot be seen.

The perpetrator suddenly runs away leaving the crowbar behind him. Some time later the Police arrives to the crime scene and seizes the crowbar. Inspecting it more in detail reveals that it has a blue colour (crowbars sold are either painted – often in red or blue – or unpainted).

In the investigation interest is taken in a certain Mr Johnson, who is a well-reputed burglar. A visit is paid at his home, but he is not there. His wife – who opened the door - is asked whether Mr Johnson is in possession of a crowbar and what it looks like. She says he has a crowbar, and it is not painted.

What do we have here?

We have a crowbar, which we know was used for the burglary attempt thanks to the CCTV take-up.

Our question is: *Is it Mr Johnson's crowbar?*

To structure things:

Let *A* denote the statement "The crowbar belongs to Mr Johnson"

Let *B* denote "The crowbar is painted in blue"

Then we have a witness' statement: *C* = "Mr Johnson's crowbar is unpainted"

How do *B* and *C* influence our belief in *A* ?

A =“The crowbar belongs to Mr Johnson”
B = “The crowbar is blue”
C = “Witness says: Mr Johnson’s crowbar is unpainted”

In terms of probabilities (using the subjective interpretation):

Why was Mr Johnson interesting from the beginning?

$P(A|I)$ must have been sufficiently high (where I is the background information available – before hearing what the witness (Mrs Johnson) said)

Is B relevant for A, i.e. is $P(A|B,I) \neq P(A|I)$?

Are A and B *conditionally dependent* given C,
i.e. is $P(A,B|C,I) \neq P(A|C,I) \cdot P(B|C,I)$?

There is a “problematic” difference between
C = “Witness says: Mr Johnson’s crowbar is unpainted”
and (what may be confused with)
C' = “Mr Johnson’s crowbar is unpainted”

A =“The crowbar belongs to Mr Johnson”
B = “The crowbar is blue”
C = “Witness says: Mr Johnson’s crowbar is unpainted”
C' = “Mr Johnson’s crowbar is unpainted”

For…

$P(A, B|C', I) = 0$ The crowbar cannot belong to Mr Johnson (A) and be blue (B) <u>if</u> Mr Johnson’s crowbar is unpainted (C')

but…

$P(A, B|C, I)$ is more difficult. In what way would the relevance between A and B be affected by a witness statement?

A =“The crowbar belongs to Mr Johnson”
B = “The crowbar is blue”
C = “Witness says: Mr Johnson’s crowbar is unpainted”
C' = “Mr Johnson’s crowbar is unpainted”

Decompose $P(A, B|C, I)$ using C' and $\neg C'$:

$$P(A, B|C, I) = P(A, B|C', C, I) \cdot P(C'|C, I) + P(A, B|\neg C', C, I) \cdot P(\neg C'|C, I) =$$

$$= 0 \cdot P(C'|C, I) + P(A, B|\neg C', C, I) \cdot P(\neg C'|C, I)$$

If $\neg C'$ holds, i.e. if Mr Johnson’s crowbar ***is*** painted, then C is no longer relevant (on its own) for A and B and we may write

$$P(A, B|C, I) = P(A|B, \neg C', I) \cdot P(B|\neg C', I) \cdot P(\neg C'|C, I)$$

$$= P(A|\neg C', I) \; ?$$

Relates to the probability that the witness is lying

Hence, since $P(A, B|C, I) \neq P(A|C, I) \cdot P(B|C, I)$ A and B cannot be considered conditionally independent given C

Belief functions

Arthur Dempster: A generalization of Bayesian Inference. *Journal of the Royal Statistical Society. Series B.* 1968, Vol. 30 (2) : 205-247.
Glenn Shafer: *Mathematical Theory of Evidence*. Princeton University Press, 1976,

constitute the grounds for *Dempster-Shafer theory of belief functions*

"(…) belief functions is a mathematical theory of how to combine degrees of rational belief derived from different evidential sources."
[Nance D. (2019). Belief functions and burdens of proof. *Law, Probability and Risk* 18: 53-76].

"The Dempster-Shafer theory, also known as the theory of belief functions, is a generalization of the Bayesian theory of subjective probability. Whereas the Bayesian theory requires probabilities for each question of interest, belief functions allow us to base degrees of belief for one question on probabilities for a related question"
[Shafer G.: *Dempster-Shafer Theory*. www.glennshafer.com/assets/downloads/articles/article48.pdf] :

Consider an event A

The axioms of probability as a measure state that $P(A) + P(\neg A) = 1$

Now, consider what is referred to as *epistemic uncertainty*.

There is evidence (knowledge) that supports belief in A ($\mathrm{supp}(A)$) to a certain amount, where support – like probability – is measured on a scale from 0 to 1.

The evidence also supports belief in $\neg A$ to a certain amount ($\mathrm{supp}(\neg A)$).

However,

$\mathrm{supp}(A) + \mathrm{supp}(\neg A)$ is not by necessity equal to 1

One may say that a portion of the total support provided by the evidence is withheld or *uncommitted* as between A and $\neg A$.

Example

Let A stand for the statement that a marketing campaign has increased the sales of a certain product in Sweden.

From marketing research it is found that the sales of the product in Stockholm in August 2023 has increased compared to August 2022, while the sales in Malmö in August 2023 has slightly decreased compared to August 2022.

The evidence (marketing research results) may then lead to that the support of A is 0.6 while the support of $\neg A$ is 0.2. Such supports could follow from considering that Stockholm has about 3 times higher population than Malmö, but these two communities cannot be said to represent fully the population of buyers in Sweden.

Hence, there is uncommitted support of 0.2 as between A and $\neg A$. This amount of support is therefore – at this stage – on the disjunction A or $\neg A$ ($\mathrm{supp}(A \vee \neg A) = 0.2$).

Support is now transformed to belief so that the belief in one single event equals the support of that event, while the belief in a disjunction is the sum of the supports of the individual events of the disjunction plus the uncommitted support (of that disjunction.

$$\mathrm{Bel}(A) = \mathrm{supp}(A) = 0{,}6$$

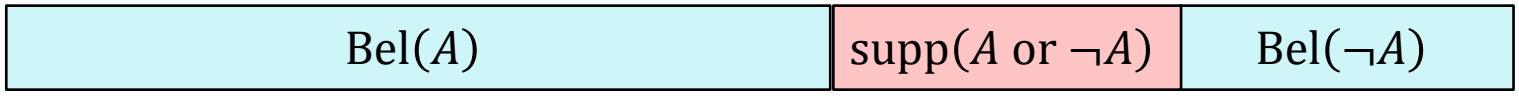
$$\mathrm{Bel}(\neg A) = \mathrm{supp}(\neg A) = 0{,}2$$

$$\mathrm{Bel}(A \text{ or } \neg A) = \mathrm{supp}(A) + \mathrm{supp}(\neg A) + \mathrm{supp}(A \vee \neg A)$$
$$= 0{,}6 + 0{,}2 + 0{,}2 = 1$$

These three values are referred to as the *belief function* .

Note that $\mathrm{Bel}(A \text{ or } \neg A) = 1 = P(A \vee \neg A)$, but $\mathrm{Bel}(A) + \mathrm{Bel}(\neg A) < 1$

The construction can be illustrated as



When the beliefs are such that $\text{Bel}(A) + \text{Bel}(\neg A) = 1$ always, the beliefs are referred to as *Bayesian beliefs*. Hence, one may say that belief functions are generalizations of subjective probabilities

Plausibility

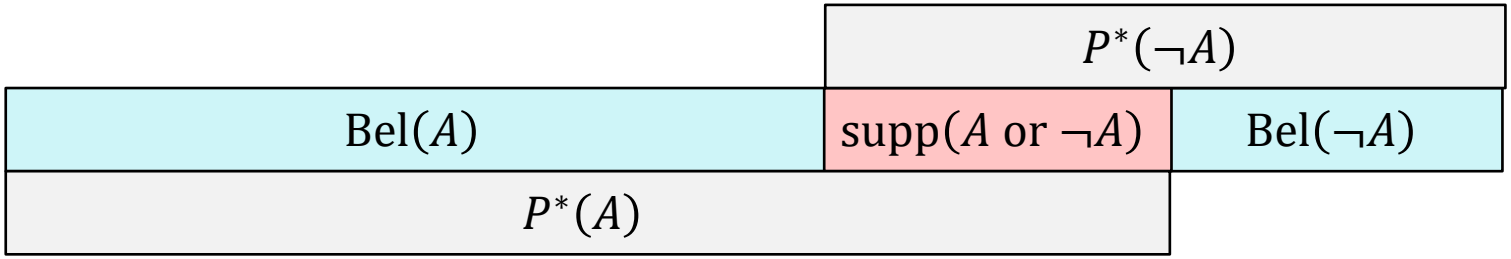
The *plausibility* (or *upper probability*) of an event A is the maximum potential belief in A:

$$P^*(A) = 1 - \text{Bel}(\neg A)$$

and the plausibility of $\neg A$ is analogously

$$P^*(\neg A) = 1 - \text{Bel}(A)$$

Graphically



Potential decision rules:

- Choose A if $BR = \dfrac{\mathrm{Bel}(A)}{\mathrm{Bel}(\neg A)} > \Lambda$
- Choose A if $PR = \dfrac{P^*(A)}{P^*(\neg A)} > \Lambda$
- Choose A if $BR_2 = \dfrac{\mathrm{Bel}(A) + \theta}{\mathrm{Bel}(\neg A) + \theta} > \Lambda$ $\qquad \theta = \dfrac{1}{2} \cdot \mathrm{supp}(A \text{ or } \neg A)$

Recall the example with the seized crowbar

A =“The seized crowbar belongs to Mr Johnson”
B = “The seized crowbar is blue”
C = “Witness says: ‘Mr Johnson’s crowbar is unpainted’”
C’ = “Mr Johnson’s crowbar is unpainted”

What is the evidence?

We know that the seized crowbar is blue and we know that the witness said that Mr Johnson’s crowbar is unpainted.

$$\Longrightarrow \text{Bel}(B) = 1, \text{Bel}(C) = 1$$

Assume we would apply decision rule 3 with $\Lambda = 1.2$
(20% exceedance of the equal stands)

$$BR_2 = \frac{\text{Bel}(A) + \theta}{\text{Bel}(\neg A) + \theta} > \Lambda$$

$$\theta = \frac{1}{2} \cdot \text{supp}(A \text{ or } \neg A)$$

$\text{supp}(A|B,C)$?

A =“The seized crowbar belongs to Mr Johnson”
B = “The seized crowbar is blue”
C = “Witness says: ‘Mr Johnson’s crowbar is unpainted’”
C’ = “Mr Johnson’s crowbar is unpainted”

We cannot forget that there are initial reasons to believe that *A* is true.
Assume $\text{supp}(A) = 0.5$.

Note that this does not imply that is $\text{supp}(\neg A)$ should also be 0.5. Assume the uncommitted support is 0.2 so that $\text{supp}(\neg A) = 0.3$. Hence, $\text{supp}(A \text{ or } \neg A) = 0.2$

Since $\text{Bel}(A) = \text{supp}(A)$ and $\text{Bel}(\neg A) = \text{supp}(\neg A)$, choosing decision rule 3 we get

$$BR_2 = \frac{\text{Bel}(A) + \theta}{\text{Bel}(\neg A) + \theta} = \frac{0.5 + 0.1}{0.3 + 0.1} = 1.5 > \Lambda = 1.2$$

$\Rightarrow$ Choose *A* !

A =“The seized crowbar belongs to Mr Johnson”
B = “The seized crowbar is blue”
C = “Witness says: ‘Mr Johnson’s crowbar is unpainted’”
C' = “Mr Johnson’s crowbar is unpainted”

Then, given B and C we might commit more support to $\neg A$ without affecting the support of A. Let’s say that we add 0.2 to the support of $\neg A$, which means that $\text{supp}(A|B,C)$ is still 0.5, while $\text{supp}(\neg A|B,C) = 0.5 \neq 0.3$

This updates the values plugged in to decision rule 3:

$$BR_2 = \frac{\text{Bel}(A|B,C) + \theta}{\text{Bel}(\neg A|B,C) + \theta} = \frac{0.5 + 0}{0.5 + 0} = 1 < \Lambda$$

$\Rightarrow$ Choose $\neg A$!

Note that we could go further analysing what would happen if we take C' into consideration, but since this is not an observable event, it cannot be used for updating.

Application to decisions of courts

Criminal law

Let H_p = "The defendant is guilty"
H_d = "The defendant is not guilty"

Depending on the country's judicial system, there may be different standards of proof.

In the Western World criminal law it is common to have
"beyond reasonable doubt"
as standard of proof

For many and historically, this means that the probability of H_p must be very high, but no common threshold is defined. Some would say 0.95, 0.98, 0.99,…

Would it work with

$$BR = \frac{\mathrm{Bel}(A)}{\mathrm{Bel}(\neg A)} > \Lambda \ ? \qquad PR = \frac{P^*(A)}{P^*(\neg A)} > \Lambda \ ? \qquad BR_2 = \frac{\mathrm{Bel}(A) + \theta}{\mathrm{Bel}(\neg A) + \theta} > \Lambda \ ?$$

Civil law

Let H_p = "The plaintiff is right"
H_d = "The respondent is right"

In the Western World civil law it is common to have as standard of proof
"preponderance of evidence" or "balance of probabilities"

For many and historically this means that the probability of H_p must be proven higher than the probability of H_d

Would it work with

$$BR = \frac{\text{Bel}(A)}{\text{Bel}(\neg A)} > \Lambda \ ? \qquad PR = \frac{P^*(A)}{P^*(\neg A)} > \Lambda \ ? \qquad BR_2 = \frac{\text{Bel}(A) + \theta}{\text{Bel}(\neg A) + \theta} > \Lambda \ ?$$