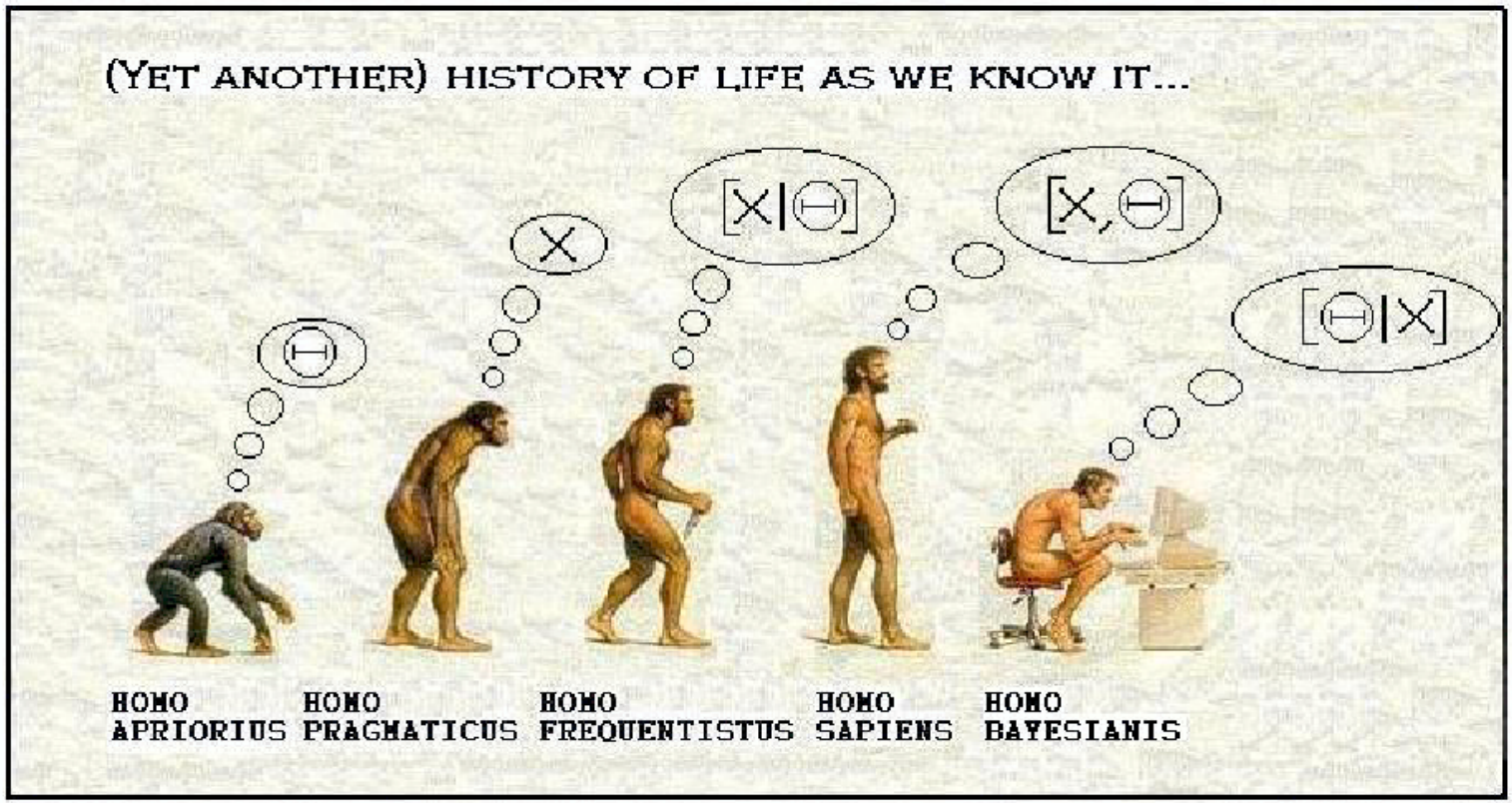
Bayesian Thinking

Dr Ana Paula Palacios

Objective

To provide an introduction to the Bayesian framework, highlighting its differences with the frequentist approach and focusing on estimation, modelisation and interpretation of the results.



Contents

Thinking Bayesian

- Suppose that when you finish your MSc you plan to submit your dissertation for publication to the *Journal of the American Statistical Association* (*JASA*).
- *JASA Theory and Methods* receives approximately 700 original submissions a year and has an acceptance rate of about 10%.
- You submit your first paper and it is accepted! What is the probability that your next submission to this journal will be accepted? Take a minute to think on this and provide a numerical answer.

Thinking Bayesian

- If your answer was greater than 10% and lower than 100%, then you are thinking Bayesian!
- From the frequentist point of view the probability that your next paper will be accepted is 100%, as the evidence suggest that your successful rate is 100%.
- But as the journal acceptance rate is only 10%, it seems reasonable to pick a number smaller than 100% and greater than 10%.

Remark I

Bayesian thinking provides a way of formalising the process of learning from data to update beliefs in accord with recent notions of knowledge synthesis.

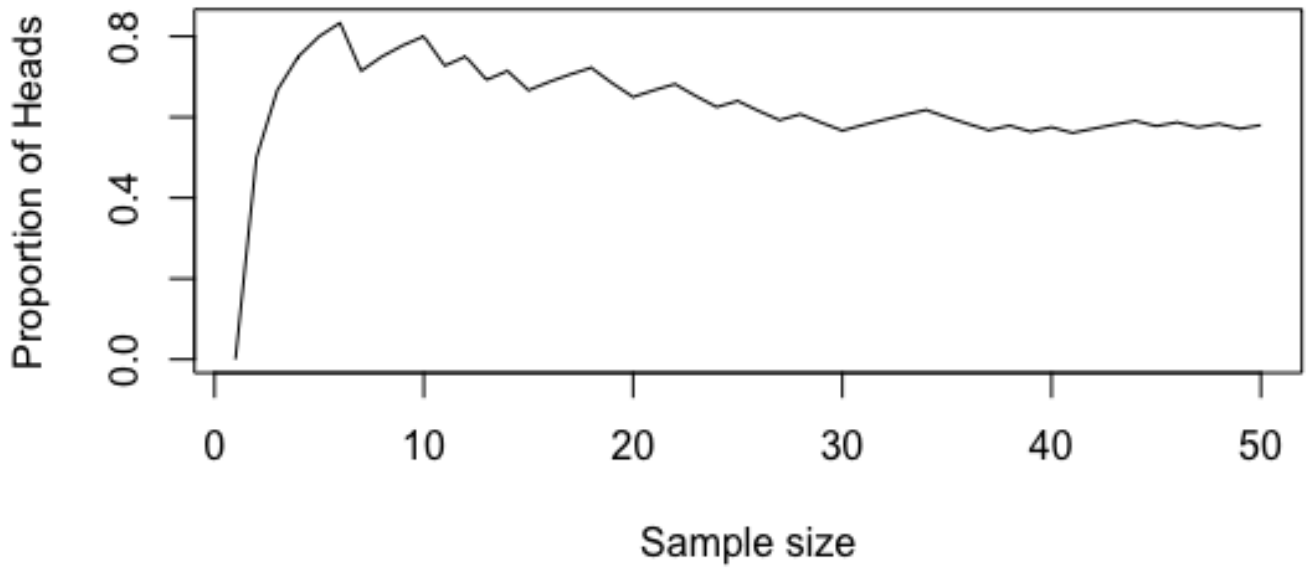
Example: coin toss

- Suppose that we have a coin and we would like to estimate the probability of Heads (θ)
- Suppose the following sequence of flips was observed: H T H H H H T H H H
- What do you think is the probability of Heads more "compatible" with this sample?
- From the frequentist point of view, the Maximum Likelihood Estimator (MLE) of the probability of Heads is $\hat{\theta} = 8/10 = 80\%$
- But, we have a lot of experience flipping coins, and my experience says that the probability of Heads should be around 0.50. Why the MLE went wrong?
- Frequentist approach relies in large sample argument: they perform well in the long-run over some infinite number of repetitions of the experiment

Probability as a Frequency

The probability of an event is equal to the long-term relative frequency of the event's occurrence when the same process is repeated many times.

Example: coin toss



Sample Size	**50**	**100**	**500**	**1000**	**5000**
$\hat{\theta}$	0.5800	0.5900	0.5180	0.5070	0.4956

Frequentist approach works well when the experiment is replicable. But, not all experiments are replicable...

Probability as uncertainty

- The Bayesian approach will include our prior beliefs.
- For example, I believe that the probability of Heads is quite likely to be around 0.50, or some value near that (0.52, 0.48) and it is less likely to be far from that (0.10, 0.90)

Bayesian Probability

Probabilities represent how certain we are about the truth of statements. These statements/beliefs can refer to the past, the present, or the future. In the most general sense, a probability is a number attached to a statement. That number specifies how likely it is that the statement is true.

Frequentists vs Bayesian

The basic philosophical difference between the frequentists and Bayesian paradigms is that

- Bayesians treat an unknown parameters θ as *random* and use probability to quantify their uncertainty about it.
- In contrast, frequentists treat θ as unknown but *fixed*, and probabilities are interpreted as long-run relative frequencies.

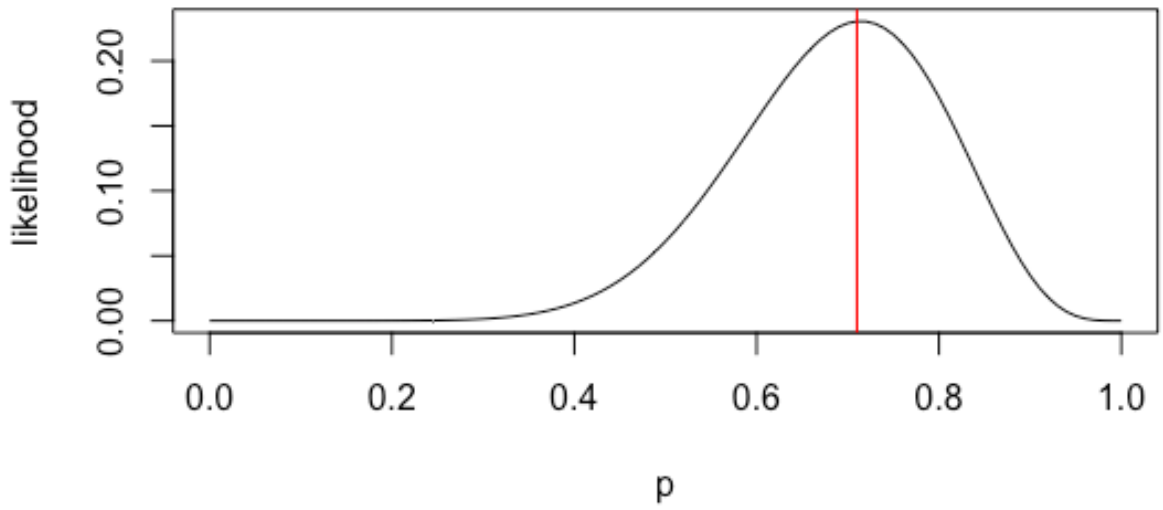
Maximum Likelihood Estimator

- Decide a model for the data (the likelihood function)
- For the coin example, the number of heads we get when flipping a coin n times follows a binomial distribution $X \sim B(n, p)$:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

  where $x = 0, 1, \ldots, n$ and $0 \leq p \leq 1$ is the probability of Heads.
- Find the value of p that maximises the likelihood function for the observed data
- Suppose we observe 10 Heads and 4 Tails ($x = 10$ and $n = 14$).

Maximum Likelihood Estimator



$$\hat{p}_{MLE} = \max_p L(x|p) = \max_p \left\{ \binom{n}{x} p^x (1-p)^{n-x} \right\}$$

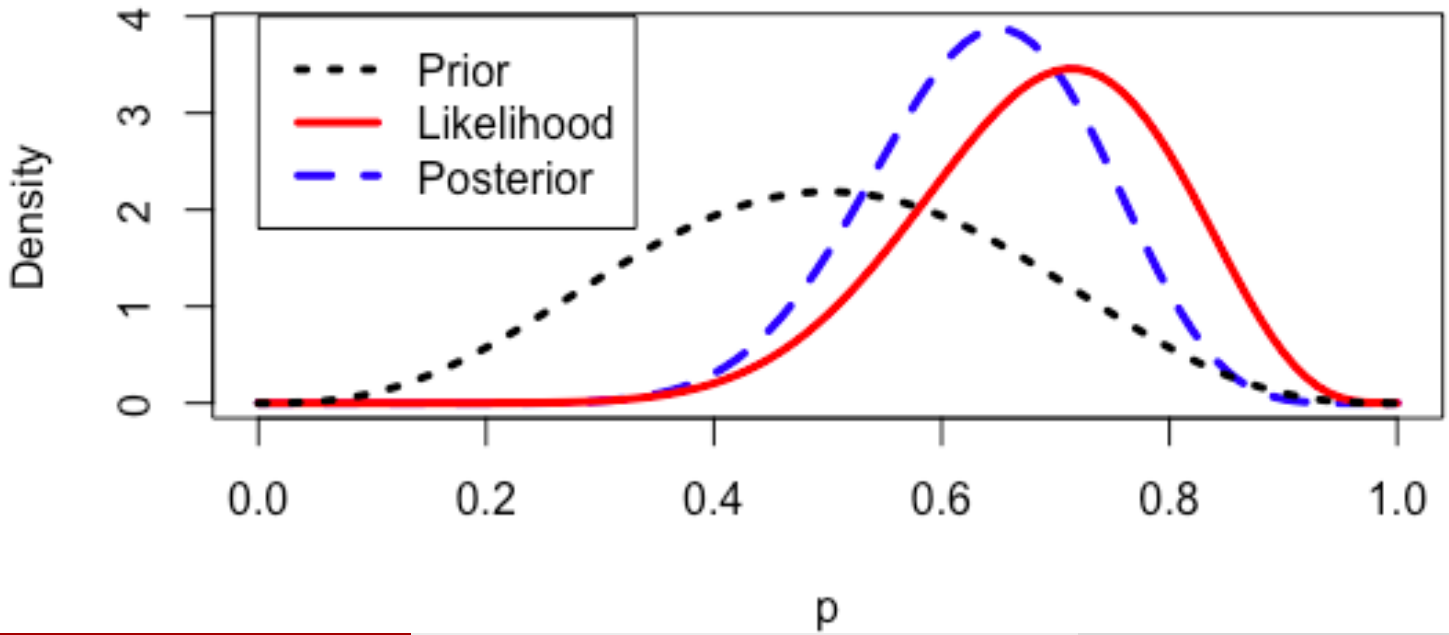
$$\hat{p}_{MLE} = \max_p \ln L(x|p) = \max_p \left\{ \ln \frac{n!}{x!(n-x)!} + x \ln p + (n-x) \ln(1-p) \right\}$$

$$\hat{p}_{MLE} = \frac{x}{n}$$

In our example: $\hat{p} = 10/14 \approx 0.71$

The Bayesian Theorem

- We are interested in the values of the unknown parameter θ
- The uncertainty about the parameter can be modelled through a probability distribution, $p(\theta)$, called the **prior distribution**. It will normally represents our *a priori* beliefs about the unknown parameter
- We have some relevant data, suppose we have n observations $x = (x_1, \ldots, x_n)$ which have a probability distribution that depends on the unknown parameter value: $p(x|\theta)$ (the likelihood).
- After observing the data, we update the beliefs about the unknown parameter.



The Bayesian Theorem

- The belief about θ are updated by applying the **Bayes Theorem** to random variables

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ p(\theta|x) &\propto p(x|\theta)p(\theta) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

where $p(x) = \int_\theta P(x|\theta)p(\theta)d\theta$. So, $p(x)$ is the normalising constant, and its distribution does not depend on the value of θ .

- The inference about θ is made is based on the distribution of θ conditional on the data, $p(\theta|x)$, called the posterior distribution

Proper posterior distribution

- To perform valid inference, the posterior distribution must be proper. For the continuous case that is $\int_\theta p(\theta|x)d\theta = 1$.
- Often we do not need to calculate the normalising constant because we can recognize the form of $p(x|\theta)p(\theta)$ as a probability distribution that we know.
- Sometimes we can derive analytically the posterior distribution
- Other times we can obtain random draws from the posterior distribution of the parameter by using Markov Chain Monte Carlo methods (Gibbs sampling, Metropolis Hasting algorithm, etc.).
- Sometimes the posterior distribution belongs to the same probability distribution family as the prior distribution

Conjugacy

If the posterior distribution belongs to the same probability distribution family as the prior distribution, the prior and posterior are then called **conjugate distributions**, and the prior is called a conjugate prior for the likelihood.

The Beta-Binomial model

Example (Bayesian Data Analysis, Gelman et al., 2003)

The proportion of births that are female has long been a topic of interest both scientifically and to the lay public. Two hundred years ago it was established that the proportion of female birth in European populations was less than 0.5, while the currently accepted value of the proportion in very large European-race populations is 0.485

- Let x be the number of girls in n recorded births. The data can be modelled with the binomial distribution, that is $X \sim B(n, \theta)$, where θ is the probability of a female birth. Therefore, the likelihood is of the form

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

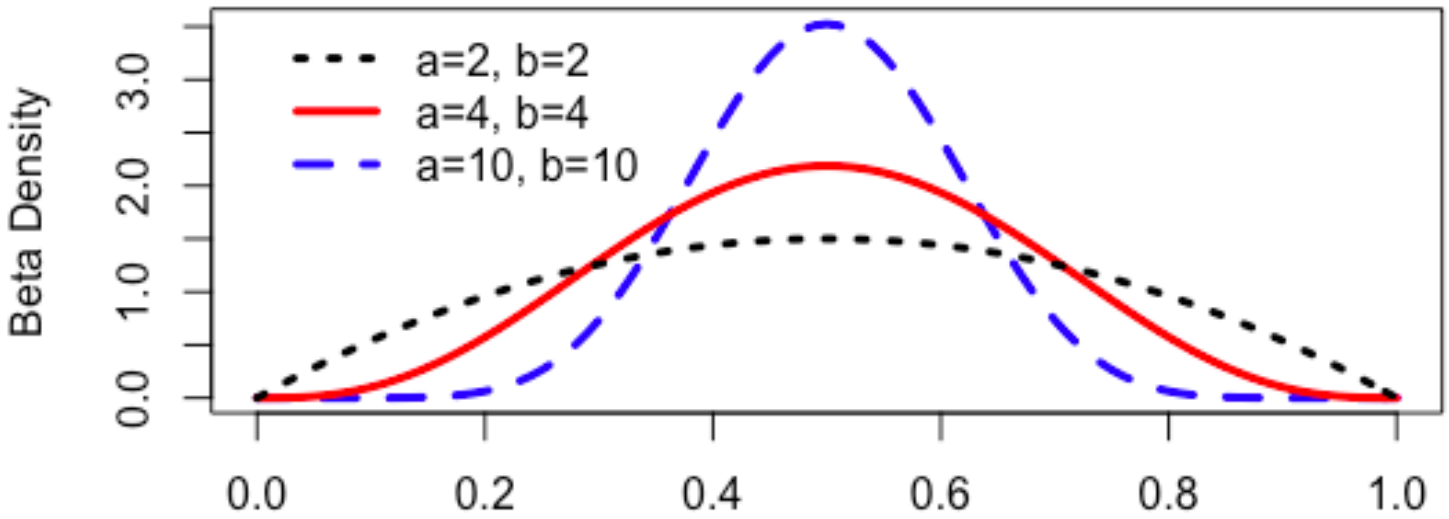
- What about the prior? Well, we know that θ is a probability and therefore $0 \leq \theta \leq 1$. It is also likely to take a value near 0.50, and less likely to be far away from 0.5. A good candidate will be the Beta distribution

The Beta-Binomial model

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{where } \theta \in [0,1]$$

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad \text{and } V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where $\alpha > 0$ and $\beta > 0$ are called hyperparameters



The Beta-Binomial model

The posterior density for θ is then

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \end{aligned}$$

Note that this looks like a Beta function with updated parameters, that is

$$P(\theta|x) = Beta(\alpha + x, \beta + n - x)$$

- As the posterior distribution follows the same parametric form as the prior distribution this is a case of conjugacy.
- Note that α and β can be seen as the prior number of success (female birth in our example) and failures (males births)

More cases of conjugacy

Likelihood	Conjuagate Prior
Binomial	Beta
Geometric	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal (σ known)	Normal
Normal (μ known)	Inverse Chi-Square
Normal (μ known)	Gamma
Exponential	Gamma

Summarising the posterior: point estimators?

How do we extract a Bayesian point estimator for the unknown parameter θ?

- The **mode** of the posterior distribution (its maximal value) is the most likely value for θ given the data:

$$\hat{\theta} = \max_{\theta}\{p(\theta|x)\}$$

- The **mean** of the posterior distribution is the posterior expected value of θ given the data

$$\hat{\theta} = E[\theta|x] = \int \theta p(\theta|x) d\theta$$

- The **median** of the posterior distribution is the value $\hat{\theta}$ that satisfies $P(\theta > \hat{\theta}|x) = P(\theta < \hat{\theta}) = 0.5$

$$\int_{\infty}^{\hat{\theta}} p(\theta|x) d\theta = \frac{1}{2}$$

Summarising the posterior: quartiles and intervals

- q-quantile of the posterior distribution: the value θ_q below which $100q$ percentage of the posterior distribution of θ fall

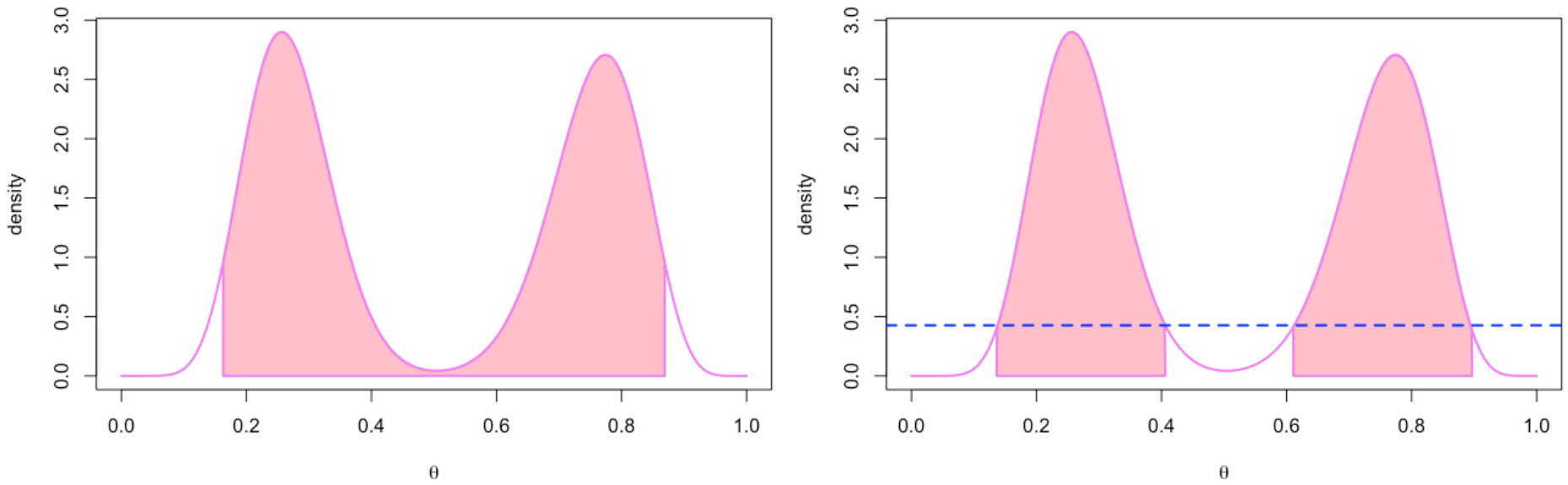
$$P(\theta < \theta_q) = q$$

- Credible interval: the $100(1-\alpha)\%$ credible interval is given by any (a, b) satisfying

$$P(a \leq \theta \leq b|x) = \int_a^b p(\theta|x)d\theta = 1 - \alpha$$

- Central posterior interval: where the lower and upper limit of the interval are the $\alpha/2$ and $1 - \alpha/2$ quantiles.
- Highest posterior density region(s): are the shortest intervals giving an area of $1 - \alpha$

Credible vs Confidence Intervals



- **Frequentist**: we are 95% confident that the true value of θ is contained in the interval. By confident we mean that if we repeat the experiment a large number of times, this will happen in 95% of the cases.
- **Bayesian**: given the observed data, there is a 95% probability that the value of θ lies in the credible interval
- A very common mistake is to do a Bayesian interpretation of a frequentist confidence interval. Remember that in the frequentist approach θ is not random.

The Predictive distribution

- Suppose now that we want to predict out-of-sample data, say $\tilde{x}$.
- We want to find the conditional distribution of $\tilde{x}$ given x.
- The posterior predictive distribution, $p(\tilde{x}|x)$ is

$$
\begin{aligned}
p(\tilde{x}|x) &= \int p(\tilde{x}, \theta|x) d\theta \\
p(\tilde{x}|x) &= \int p(\tilde{x}|\theta, x) p(\theta|x) d\theta \\
p(\tilde{x}|x) &= \int p(\tilde{x}|\theta) p(\theta|x) d\theta
\end{aligned}
$$

as $\tilde{x}$ and x are conditionally independent given θ.

The Predictive distribution

Continuing with the Beta-Binomial model, the posterior predictive distribution is:

$$
\begin{aligned}
p(\tilde{x}|x) &= \int p(\tilde{x}|\theta)p(\theta|x)d\theta \\
&= \int_0^1 \binom{m}{\tilde{x}} \theta^{\tilde{x}}(1-\theta)^{m-\tilde{x}} \frac{1}{B(\alpha^*,\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}d\theta \\
&= \binom{m}{\tilde{x}} \frac{1}{B(\alpha^*,\beta^*)} \int_0^1 \theta^{\tilde{x}}(1-\theta)^{m-\tilde{x}}\,\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}d\theta \\
&= \binom{m}{\tilde{x}} \frac{1}{B(\alpha^*,\beta^*)} \int_0^1 \theta^{\tilde{x}+\alpha^*-1}(1-\theta)^{m-\tilde{x}+\beta^*-1}\,d\theta \\
&= \binom{m}{\tilde{x}} \frac{1}{B(\alpha^*,\beta^*)} B(\tilde{x}+\alpha^*, m-\tilde{x}+\beta^*)\,d\theta \\
\tilde{x}|x &\sim \text{Beta-bin}(m,\alpha^*,\beta^*)
\end{aligned}
$$

where $\alpha^* = \alpha + x$ and $\beta^* = \beta + n - x$.

Choosing a Prior distribution

- The relative influence of the prior and data on updated beliefs depends on how much weight is given to the prior knowledge and the strength of the data.
- If the data has sufficient signal, then the prior will not influence greatly the posterior
- But if the data is weak or the sample size is small, then the choice of the prior may have a big impact in the posterior
- Normally, the location of a parameter (mean, mode) and its precision of the prior is usually more critical than the shape of the distribution
- If some prior information is available (past historical data, expert opinion, etc), one should choose a prior distribution that reflects this prior knowledge. This will be particularly useful when the sample size is small or the parameter space is of high dimension

Informative vs Non-informative priors

Informative-Subjective priors

They represent our prior beliefs about parameter values before collecting any data. The process of constructing the most suitable prior for θ is called prior elicitation.

Weakly informative priors

They express partial information about the parameter. They usually have very high variance.

Noninformative-Objective Priors

They contain little or none information about θ (they do not favour any value of θ). They are usually improper. The posterior distribution will highly depend on the likelihood and inferences in this case are called objective. Examples:

- $U(-\infty, \infty)$
- $B(0, 0)$
- Jeffreys prior distribution

Constructing a prior

- Returning to the example of female births, we choose to have a beta prior distribution. How can we choose the values of the hyperparameters α and β?
- We can interpret α and β as previous "successes" and "failures"
- Experts may have some idea about the mean value, the variance, some percentile, etc. We can use this information to derive the corresponding value of α and β.
- For example, if we have some knowledge about the mean and the variance, then we can solve a system of equations to find out the value of α and β.

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad \text{and}$$

$$V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Summarising

Frecuentists	Bayesians
θ is fixed	θ is a random variable
Data is random (and replicable)	Data is fixed (given)
Probability as frequency	Probability as uncertainty
Prior knowledge is ignored	Prior knowledge is included
Confidence intervals	Credible intervals
Hypothesis test	Posterior distribution of θ