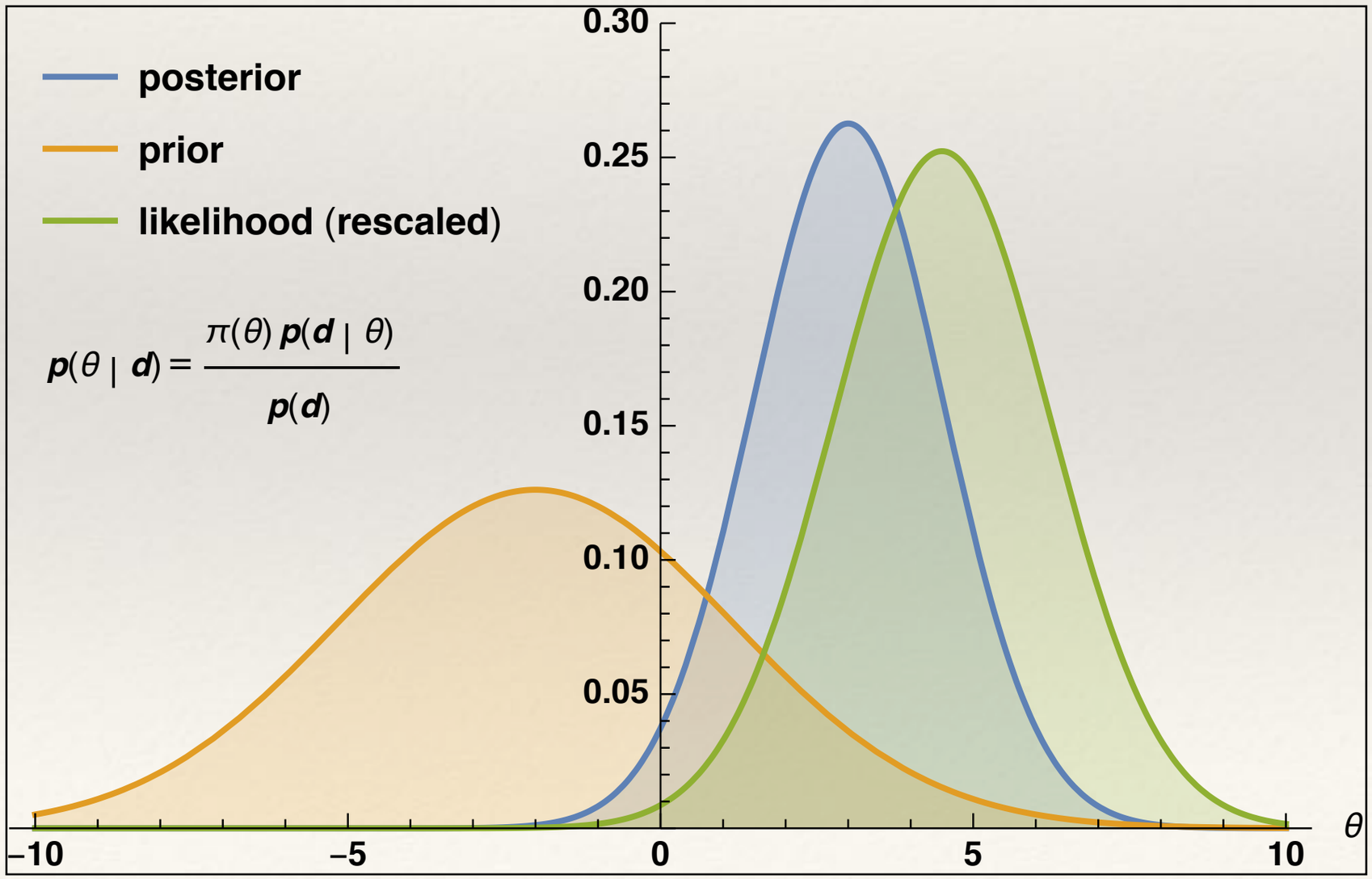
Making sense of data: introduction to statistics for gravitational wave astronomy

Lecture 4: Bayesian inference part I

AEI IMPRS Lecture Course

Jonathan Gair jgair@aei.mpg.de

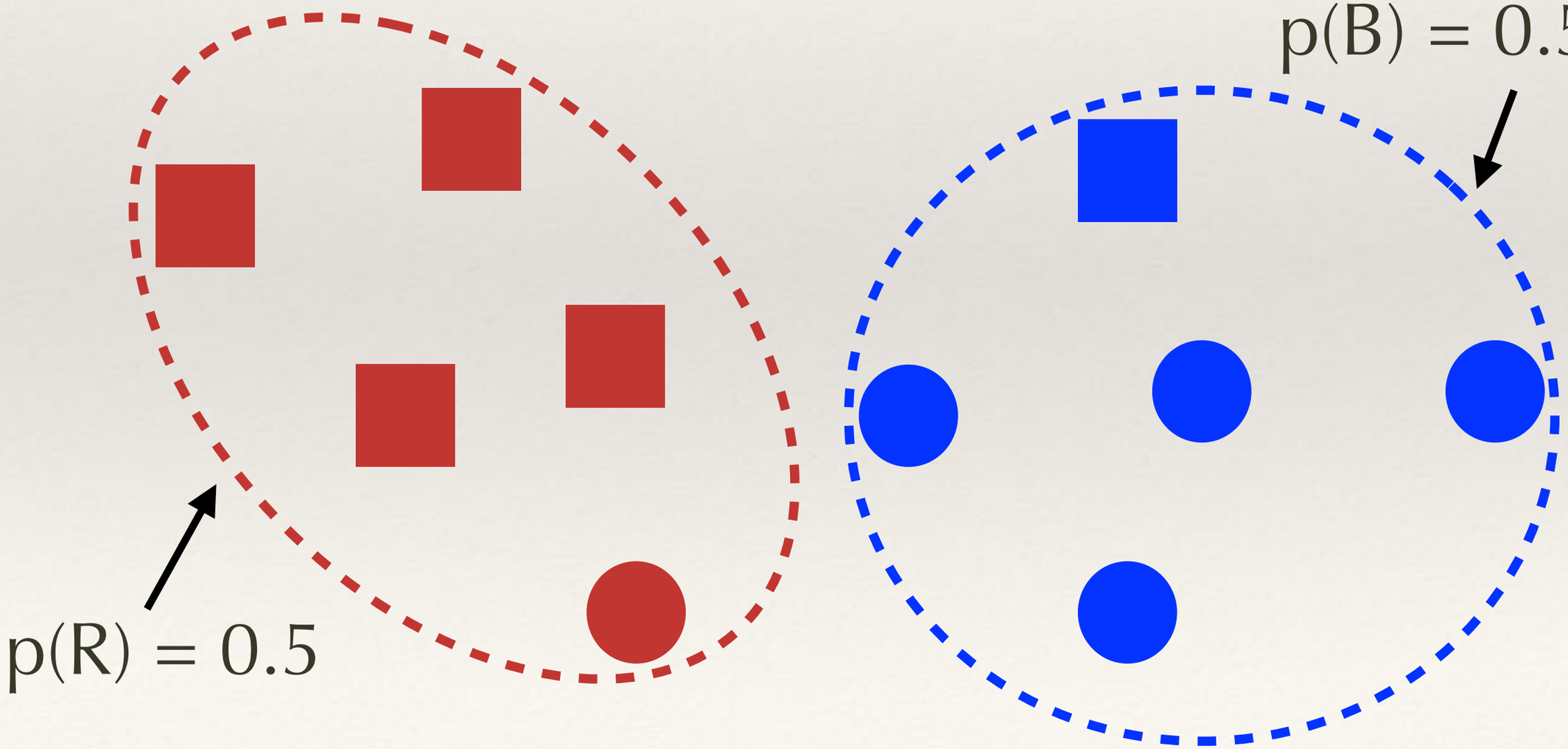


Bayesian versus Frequentist Statistics

- Frequentist statistics makes references to repeated experiments with parameters fixed but unknown.
- In Bayesian inference the parameter values are regarded as random variables.
- In a given observation, the pdf of the parameters is updated from a **prior** to a **posterior** using the likelihood of the observed data.
- A Bayesian posterior can be interpreted as a probability distribution on the parameter values based on the observed data set. It is based only on the observed data and does not make reference to hypothetical repetitions of the experiment.
- In a GW context, Bayesian inference makes intuitive sense as experiments are not repeatable. Even when they are, the Bayesian posterior converges to the fixed but unknown parameter value as the number of experiments increases.
- Bayes' Theorem is a mathematical identity. The distinction between frequentist and Bayesian inference is philosophical, in the interpretation of various quantities.

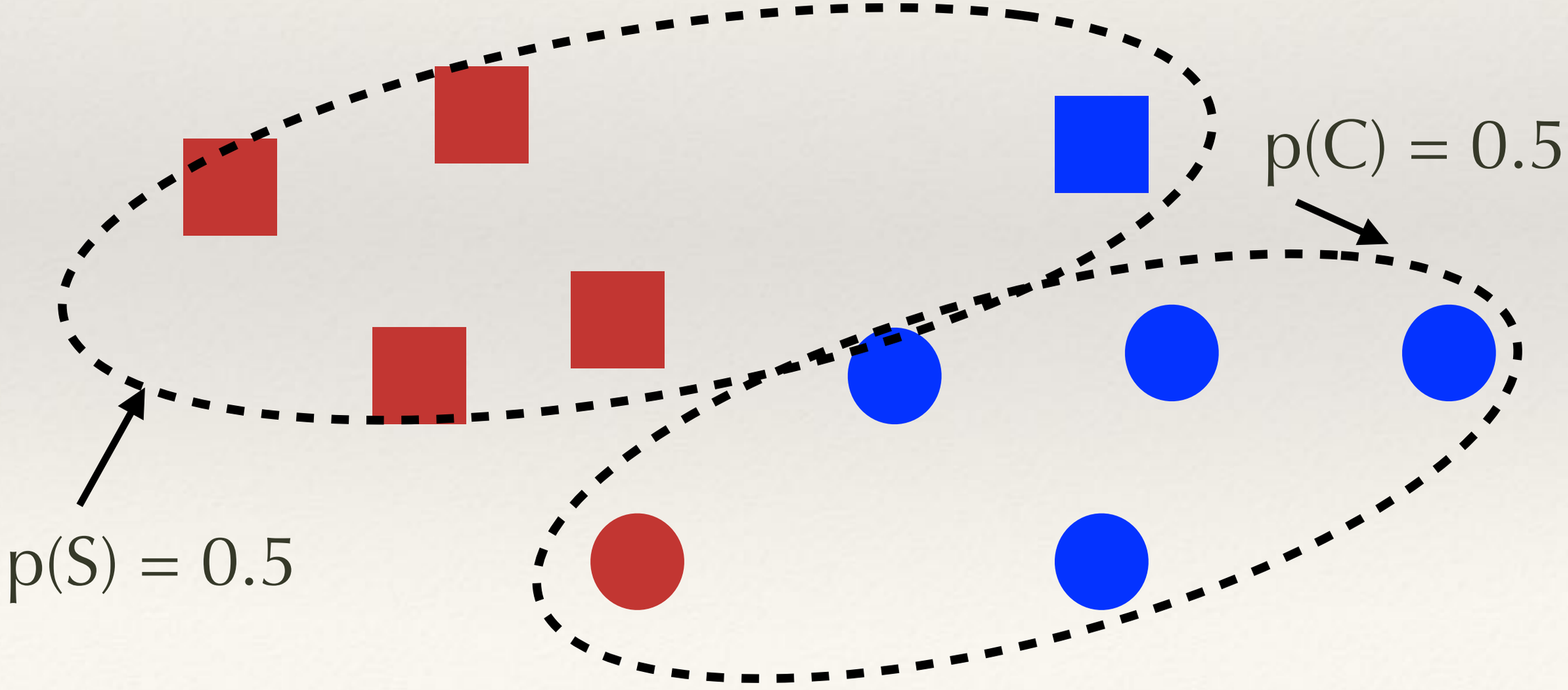
Conditional probability

- Suppose we choose at random from a set of objects that are red (R) or blue (B) and circular (C) or square (S).
- 5 out of 10 objects are red, therefore, with no other information, $p(R)=0.5$.



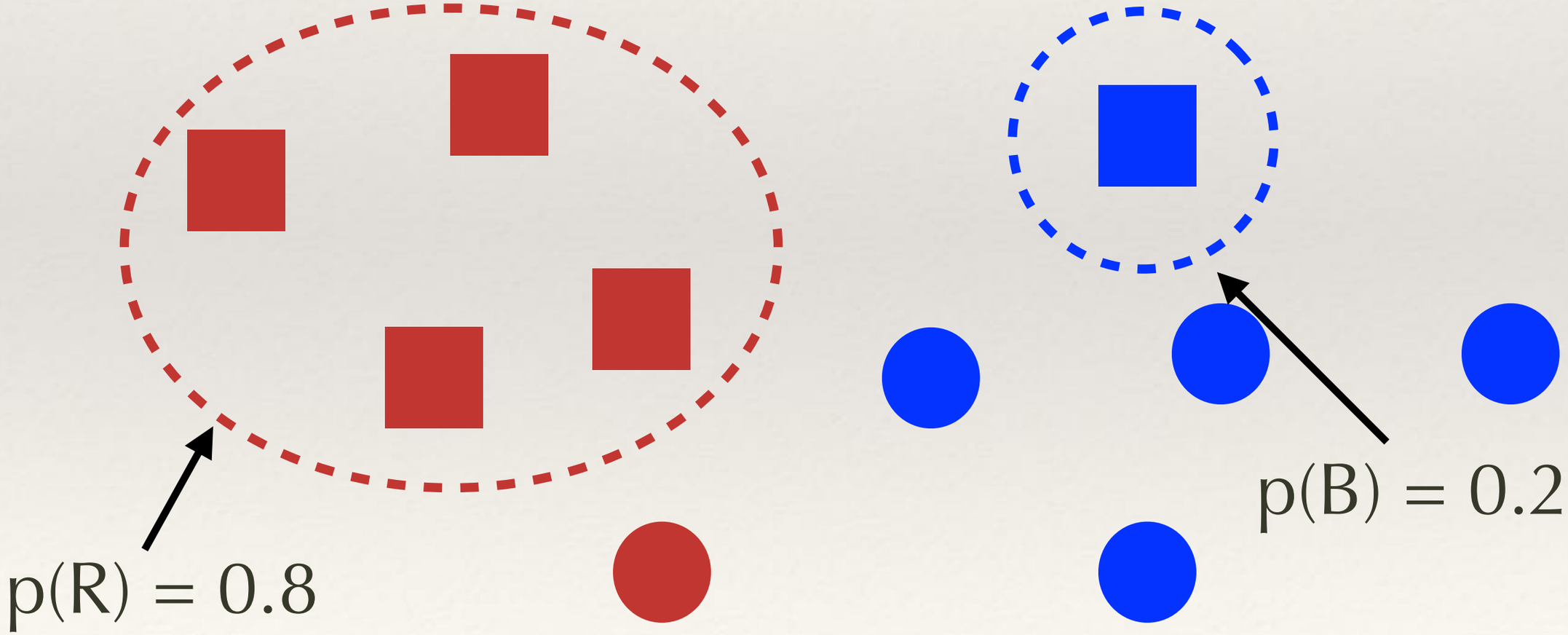
Conditional probability

- Suppose we choose at random from a set of objects that are red (R) or blue (B) and circular (C) or square (S).
- 5 out of 10 objects are red, therefore, with no other information, $p(R)=0.5$.



Conditional probability

- If we know that the object is square, the probability changes, since of square objects, 4 out of 5 are blue.



Conditional probability

- The idea that, for correlated random variables, the distribution of one can change based on the observed value of the other is encoded in the notion of *conditional probability*.
- Mathematically we define the probability of A *given* B as

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- If A and B are independent then $p(A \mid B) = p(A)$.

Conditional probability

- The idea that, for correlated random variables, the distribution of one can change based on the observed value of the other is encoded in the notion of *conditional probability*.
- Mathematically we define the probability of A *given* B as

Probability that A and B occur simultaneously

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- If A and B are independent then $p(A \mid B) = p(A)$.

Conditional probability

- The idea that, for correlated random variables, the distribution of one can change based on the observed value of the other is encoded in the notion of *conditional probability*.
- Mathematically we define the probability of A *given* B as

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Probability that B occurs

- If A and B are independent then $p(A \mid B) = p(A)$.

Bayes' Theorem

- Rearranging the definition of conditional probability

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- we obtain Bayes' Theorem:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- This is mathematically exact, but can be used in an approximate way for inference
  - p(A) — prior belief about state of the Universe, "A";
  - p(B | A) — likelihood of seeing data "B" if the state is "A";
  - p(A | B) — posterior belief on the state of the Universe after collecting data;
  - p(B) — "evidence" for your model (a normalising constant).

Bayesian inference: example

- Suppose a medical screening test is 95% **effective** but has a 1% **false alarm rate**. This means that the probability of getting a positive result when the patient *does have* the disease is 0.95, while the probability of getting a positive result when the patient *does not* have the disease is 0.01.

- If the disease has a prevalence of 0.5% in the population, what is the probability that a person with a positive result has the disease?

$$p(\text{ill}|\text{pos}) = \frac{p(\text{pos}|\text{ill})p(\text{ill})}{p(\text{pos}|\text{ill})p(\text{ill}) + p(\text{pos}|\text{well})p(\text{well})}$$

$$= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995}$$

$$= \frac{0.00475}{0.00475 + 0.00995} = 0.323$$

Prior choice: informative

- The prior plays a key role in Bayesian inference. One advantage of the Bayesian approach is the ability to include additional information through the prior.
- Priors can be **informative** or **uninformative**. An informative prior makes a (strong) statement about the values or distribution of the parameters under consideration. Uninformative priors attempt to say as little as possible and thereby avoid biasing the results.
- Informative priors may come from previous experiments, i.e., these could be the posterior from a previous (set of) experiment(s). Alternatively they can be based on the opinion of "experts", through their experience in similar situations.
- The process of constructing a prior based on expert input is known as **elicitation**.
- Different experts may have different opinions, in which case **mixture priors** can be used

$$p(\vec{\theta}) = \sum_{j=1}^{J} \omega_j p_j(\vec{\theta})$$

Prior choice: conjugate priors

- Another commonly used approach to prior definition is to use **conjugate priors**.

Definition: A family of distributions, $\mathcal{F}$, is **conjugate** to a family of sampling distributions, $\mathcal{P}$, if, whenever the prior belongs to the family $\mathcal{F}$, the posterior belongs to the same family, for any number and value of observations from $\mathcal{P}$.

- In other words, the posterior is from the same family as the prior and can therefore be used as the prior for the next observation and so on. Any distribution in the exponential family

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^{K} A_j(x)B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\}\ \forall x, \vec{\theta}$$

- has a conjugate prior of the form

$$p(\vec{\theta}|\vec{\chi}, \nu) = p(\vec{\chi}, \nu)\exp\left[\vec{\theta}^T\vec{\chi} - \nu A(\vec{\theta})\right]$$

- The most commonly encountered conjugate prior models are Beta-Binomial, Poisson-Gamma and Normal-Normal/Normal-Gamma.

Conjugate priors: Beta-Binomial

- For binomial observations with likelihood

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

- the conjugate prior is the Beta distribution.

$$p(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$$

- The posterior is then

$$\begin{aligned} p(p \mid x) &\propto p(x \mid p)p(p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \\ &\propto p^{a+x-1}(1-p)^{b+n-x-1} = \text{Beta}(a+x, b+n-x) \end{aligned}$$

- The mean of a *Beta(a,b)* distribution is *a/(a+b)* so the posterior mean is

$$\mathbb{E}(p|x) = \frac{a+x}{a+b+n}$$

Conjugate priors: Poisson-Gamma

- For Poisson-distributed observations

$$p(\mathbf{x} \mid \lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\}$$

- the conjugate prior is a Gamma distribution

$$p(\lambda | m, \mu) = \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda}$$

- The posterior is then

$$\begin{aligned} p(\lambda \mid \mathbf{x}) &\propto p(\mathbf{x}|\lambda) p(\lambda) \\ &= \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\} \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda} \\ &\propto \mathrm{e}^{-n\lambda-\mu\lambda} \lambda^{\sum_{i=1}^{n} x_i + m - 1} \\ &\propto \mathrm{Gamma}(m + n\bar{x}, \mu + n). \end{aligned}$$

- for which the posterior mean is

$$\mathbb{E}(p(\lambda \mid \mathbf{x})) = \frac{m + n\bar{x}}{m + n} = \bar{x} \left(\frac{n}{n + m} \right) + \frac{m}{\mu} \left(1 - \frac{n}{n + m} \right)$$

Conjugate priors: Normal-Normal

- We now consider $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with likelihood

$$p(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- If we assume the variance is known the conjugate prior is also Normal

$$p(\mu \mid \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right]$$

- The posterior

$$\begin{aligned} p(\mu \mid \mathbf{x}, \sigma^2) &\propto p(\mathbf{x} \mid \mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2\sigma_0^2}\left[\mu^2(n\sigma_0^2 + \sigma^2) - 2\mu(n\bar{y}\sigma_0^2 + \mu_0\sigma^2)\right]\right\} \end{aligned}$$

- which can be recognised as a Normal distribution with mean and variance

$$\mu_n = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

Conjugate priors: Normal-Gamma

- If we now assume instead that the mean is known, but the variance is not, the conjugate prior on the **precision** (the reciprocal of the variance) is a Gamma distribution

$$p(\tau|a,b) \propto \tau^{a-1}e^{-b\tau}$$

- with posterior $p(\tau \mid \mathbf{x}, \mu) \propto p(\mathbf{x} \mid \mu, \tau)p(\tau|a,b)$

$$\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\tau^{a-1}e^{-b\tau}$$

$$= \tau^{a+n/2-1} \exp\left\{-\tau\left(b+\frac{1}{2}\sum_{i}(x_i-\mu)^2\right)\right\}$$

$$\sim \text{Gamma}\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

- Common practice is to take a and $b << 1$ and then the posterior is approximately

$$p(\tau \mid \mathbf{x}, \mu) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \quad \Rightarrow \quad \mathbb{E}\left[\tau \mid \mathbf{x}, \mu\right] = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-1}$$

Conjugate priors: Normal-Gamma

- The final variant of the Normal-Normal model is to assume both mean and variance are unknown. In that case a conjugate prior can be found of the form

$$\mu \sim N(\mu_0, 1/(n_0\tau)), \quad \tau \sim \text{Gamma}(a, b)$$

- The posterior on the mean is the same as before

$$p(\mu|\tau, \mathbf{x}) \sim N\left(\frac{n_0\mu_0 + n\bar{x}}{n_0 + n}, \frac{1}{(n_0 + n)\tau}\right)$$

- but the posterior on the precision is now

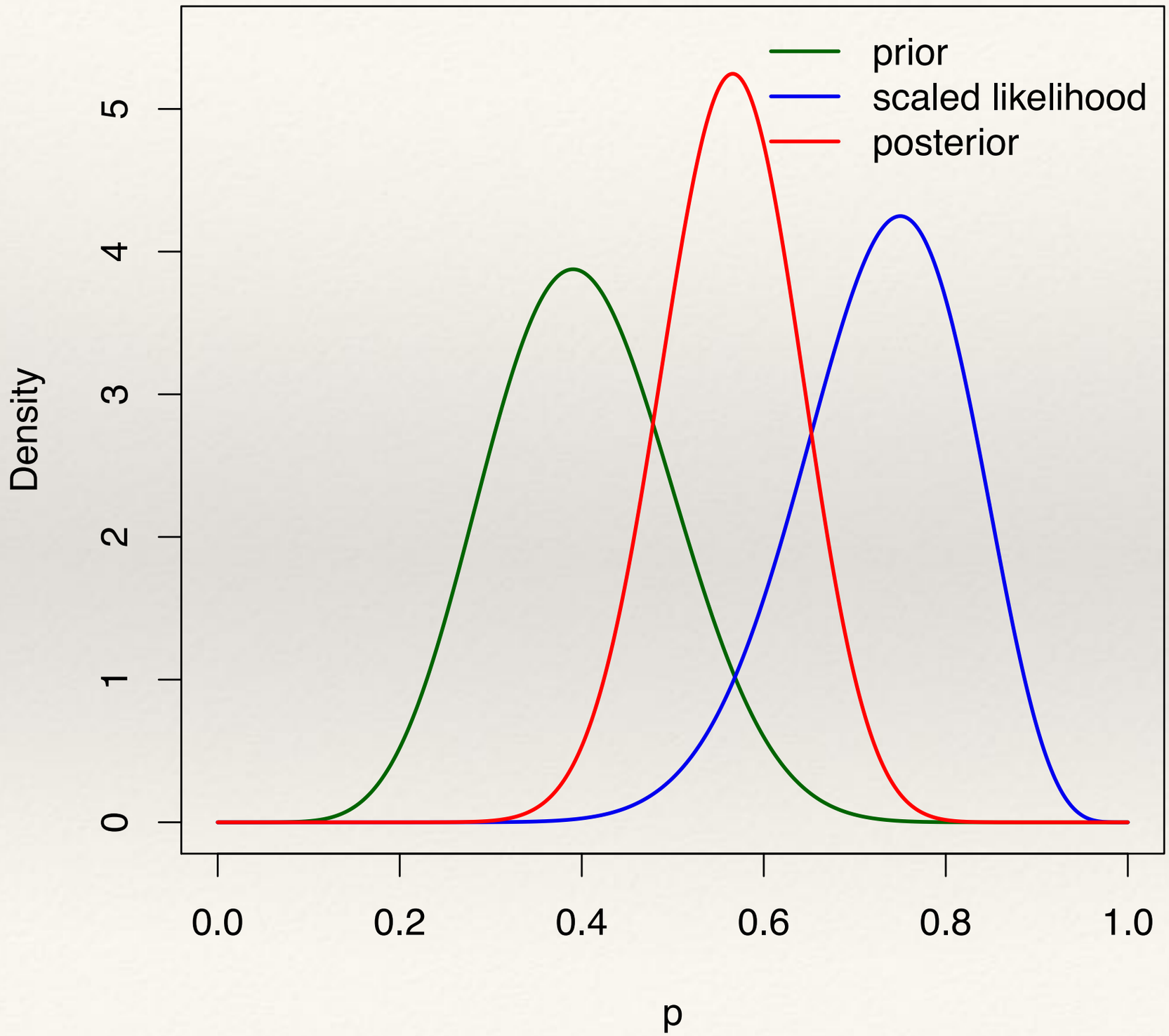
$$p(\tau|\mathbf{x}) \sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_0}{2(n + n_0)}(\mu_0 - \bar{x})^2\right)$$

Using expert information in conjugate priors

- Informative priors that are also conjugate can be constructed, if the form in which the the expert information is available is appropriate.
- **Example**: *Consider a drug to be given for relief of chronic pain. Experience with similar compounds has suggested that response rates, p, between 0.2 and 0.6 could be feasible. We plan to observe the response rate in n patients and want to infer a posterior on p. Propose a suitable conjugate prior for p based on the available information. Hence obtain the posterior if x = 15 positive responses are observed in a sample of n=20 patients.*

    - The expert information can be interpreted as *U[0.2,0.6]*.
    - A *U[0.2,0.6]* distribution has mean 0.4 and variance 0.01.
    - These are the same mean and variance as a *Beta(9.2, 13.8)* distribution. Therefore we use this is the conjugate prior.
    - For *n=20, x=15* the posterior is *Beta(24.2, 18.8)*.

Using expert information in conjugate priors



Prior choice: Jeffreys prior

- In the absence of any previous information it is desirable to choose an **uninformative** prior. Uniform priors are often regarded as uninformative. However, these are not invariant under transformations.

- Jeffreys (1961) proposed an invariant prior of the form

$$p(\vec{\theta}) \propto \sqrt{\det[I(\vec{\theta})]}, \qquad \text{where } I(\vec{\theta})_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right]$$

- is the Fisher Information matrix.

- **Example - Poisson distribution:** For the likelihood of a single Poisson observation

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

- we have

$$\frac{\partial \log p}{\partial \lambda} = \frac{x}{\lambda} - 1, \quad \frac{\partial^2 \log p}{\partial \lambda^2} = -\frac{x}{\lambda^2} \quad \Rightarrow \quad I(\lambda) \equiv \mathbb{E}\left[-\frac{\partial^2 \log p}{\partial \lambda^2}\right] = \frac{1}{\lambda}$$

- This is the prior used in standard (FGMC) LVC rate estimation.

Posterior summaries: point estimates

- The output of Bayesian inference is a probability distribution. It is often convenient to summarise the posterior in various ways, as discussed in Lecture 1.
- Usually summaries are computed for individual parameters using the **marginal distributions**

$$p_{\mathrm{marg}}(\theta_1|\mathbf{x}) = \int p(\vec{\theta}|\mathbf{x})\mathrm{d}\theta_2 \ldots \mathrm{d}\theta_m$$

- The **posterior mean** is defined by

$$\mu = \int_{-\infty}^{\infty} \theta_1 p_{\mathrm{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1$$

- The **posterior median** m is defined through

$$\int_{-\infty}^{m} p_{\mathrm{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = 0.5 = \int_{m}^{\infty} p_{\mathrm{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1$$

- The **posterior mode** is given by

$$M = \mathrm{argmax}\ p_{\mathrm{marg}}(\theta_1|\mathbf{x})$$

- The mean and mode are also well defined for the full (non-marginal) distribution.

Posterior summaries: intervals

- The Bayesian analogue of a frequentist confidence interval is a **credible interval**

  Definition: An interval (a, b) is a $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_a^b p_{\text{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = (1-\alpha), \quad 0 \leq \alpha \leq 1.$$

- A **credible region** can be defined analogously.

- Credible regions/intervals are not unique. The two most common types of credible intervals are **symmetric** and **highest posterior density** intervals.

  Definition: An interval (a, b) is a **symmetric** $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_{-\infty}^a p_{\text{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = \frac{\alpha}{2} = \int_b^{\infty} p_{\text{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1.$$

  Definition: An interval (a, b) is a $100(1-\alpha)\%$ **highest posterior density (HPD) interval** for θ_1 if

  1. $[a, b]$ is a $100(1-\alpha)\%$ credible interval for θ_1;

  2. for all $\theta \in [a, b]$ and $\theta' \notin [a, b]$ we have $p_{\text{marg}}(\theta|\mathbf{x}) \geq p_{\text{marg}}(\theta'|\mathbf{x})$.

Posterior summaries: samples

- Constructing summary statistics throws away information that can only be captured by the full posterior distribution. Sometimes the posterior distribution is expressible in closed form, but usually it is not.

- The majority of applications of probability distributions reduce to computing integrals. Another way to summarise a posterior is therefore to generate a large number of samples $\{\vec{\theta}_1, \ldots, \vec{\theta}_M\}$ from the posterior. Posterior integrals can then be computed via

$$\int f(\vec{\theta}) p(\vec{\theta}|\mathbf{x}) \mathrm{d}\vec{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} f(\vec{\theta}_i)$$

- In Lecture 6 we will discuss various methods through which such samples can be generated efficiently in practice.

Posterior summaries: interpretation

- Posterior summary statistics can be interpreted using **decision theory**. Central to decision theory is the notion of a **loss function** and the associated **risk**

$$R(\theta, d) = \mathbb{E}_\theta L(\theta, d(X)) = \begin{cases} \sum_{x \in \mathcal{X}} L(\theta, d(x)) p(x; \theta) & \text{for discrete } X \\ \int_{\mathcal{X}} L(\theta, d(x)) p(x; \theta) \mathrm{d}x & \text{for continuous } \mathcal{X} \end{cases}$$

- The **Bayes risk** of a decision rule is the expected risk with respect to the prior

$$r(\pi, d) = \int_{\theta \in \Omega_\theta} R(\theta, d) \pi(\theta) \mathrm{d}\theta$$

- A **Bayes rule** minimises the Bayes risk and can also be seen to minimise the **posterior expected loss**

$$\begin{aligned} r(\pi, d) &= \int_{\Omega_\theta} R(\theta, d) \pi(\theta) \mathrm{d}\theta \\ &= \int_{\Omega_\theta} \int_{\mathcal{X}} L(\theta, d(x)) p(x|\theta) \pi(\theta) \mathrm{d}x \mathrm{d}\theta \\ &= \int_{\Omega_\theta} \int_{\mathcal{X}} L(\theta, d(x)) p(\theta|x) p(x) \mathrm{d}x \mathrm{d}\theta \\ &= \int_{\mathcal{X}} p(x) \left\{ \int_{\Omega_\theta} L(\theta, d(x)) p(\theta|x) \mathrm{d}\theta \right\} \mathrm{d}x \end{aligned}$$

Posterior summaries: interpretation

- The Bayes rule for different loss functions corresponds to various natural summary statistics.

  - **Squared error loss**: use **posterior mean**.

$$L(\theta, d) = (\theta - d)^2$$

  - **Absolute magnitude loss**: use **posterior median**.

$$L(\theta, d) = |\bar{\theta} - d|$$

  - **Delta-function gain**: use **posterior mode**.

$$L(\theta, d) = \begin{cases} -\delta(\theta - d) & \text{if } d = \theta \\ 0 & \text{if } d \neq \theta \end{cases}$$

  - **Interval estimation**: for the loss function below, the Bayes rule is the **highest posterior density interval**.

$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta \\ 1 & \text{if } |\theta - d| > \delta \end{cases}$$

Example: linear model

- Suppose we have data

$$y_i \sim N(\mathbf{x}_i^T \vec{\beta}, \sigma^2), \quad i = 1, \ldots, N$$

- We want to fit a Bayesian model for the unknown parameters in the model. We first need to specify priors

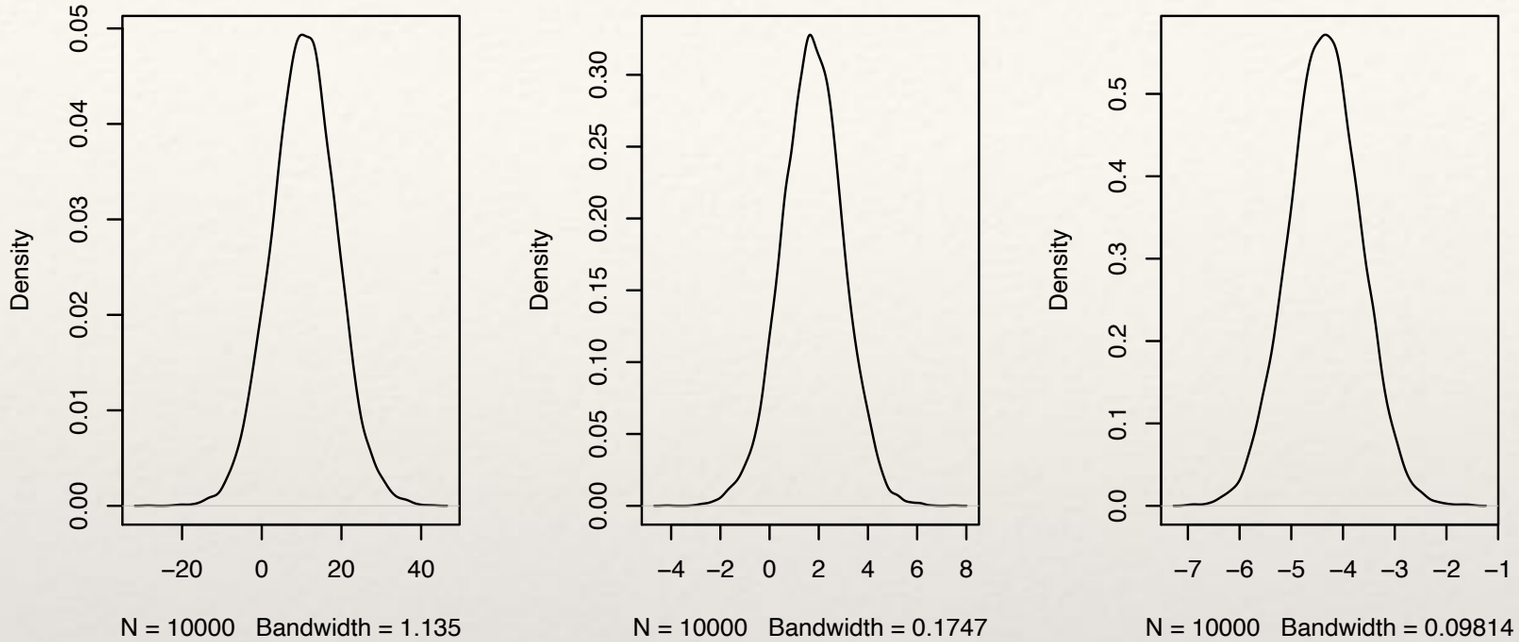
$$p(\vec{\beta}, \tau) = p(\tau) \prod_{i=1}^{p} p(\beta_j)$$

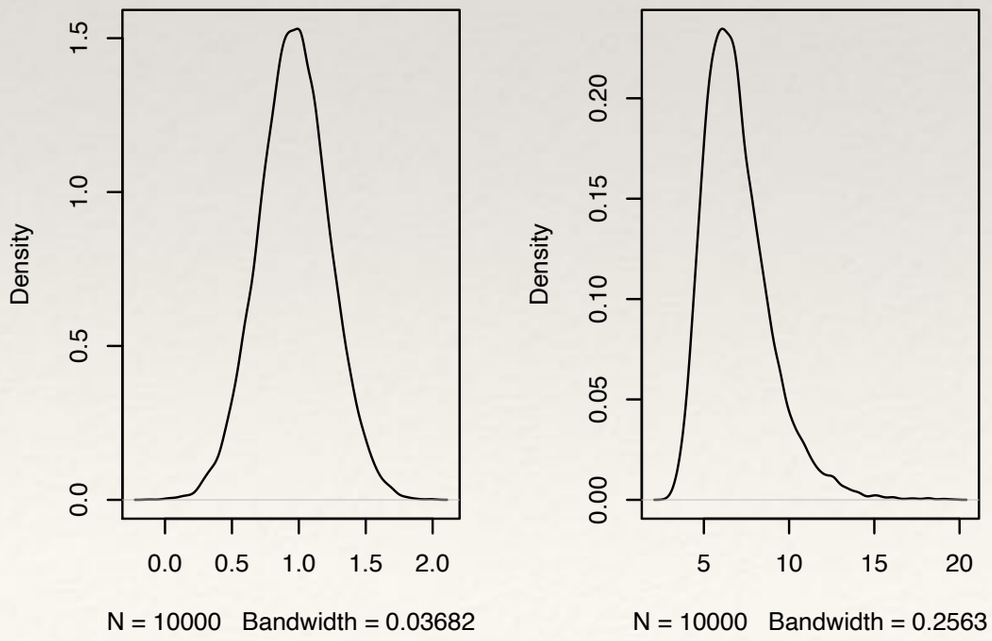
$$\beta_j \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j}), \quad \tau \sim \text{Gamma}(a, b)$$

- We use "skeptical priors" by setting the means of the Normal distribution to zero, the variances to 1000 and $a = b = 0.1$.

- We fit the standard mtcars data set, which has 3 explanatory variables and N=32 measurements.

Example: linear model





Example: linear model

- We can compare the Bayesian inference results to the maximum likelihood estimator.

	Bayesian results		Frequentist results	
Parameter	Posterior mean	95% credible interval	MLE	95% confidence interval
β_0	10.369	[-5.098,36.349]	11.395	[-5.134,27.922]
β_1	1.777	[-0.721,4.166]	1.750	[-0.857,4.169]
β_2	-4.335	[-5.702,-2.995]	-4.347	[-5.787,-3.009]
β_3	0.968	[0.449,1.493]	0.946	[0.410,1.482]
σ^2	6.978	[4.160,11.729]	6.554	—

Example: linear model

- We can diagnose the quality of the model by looking at *studentised residuals*.

$$\hat{\epsilon}_i = \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$