Importance sampling

We are interested in

$$\mu = E_f(h(x)) = \int h(x)f(x)dx$$

- If possible compute it analytically!
- If we can sample from $f(x)$ we can use Monte Carlo integration
- Possible alternative: Importance sampling
  - sample from ausiliary distribution $g(x)$ and re-weight
  - can be used as variance-reduction technique

Importance sampling Algorithm

Let $x_1, \ldots, x_n \sim g(x)$, and let $w(x_i) = \frac{f(x_i)}{g(x_i)}$, $i = 1, \ldots, n$ then

$$\hat{\mu}_{IS} = \frac{\sum h(x_i) w(x_i)}{n}$$

- Unbiased
- Consistent
- Need to know the normalizing constant

$$\tilde{\mu}_{IS} = \frac{\sum h(x_i) w(x_i)}{\sum w(x_i)}$$

- Biased for finite n
- Consistent
- Self-normalizing

Bayesian concept

. . . The essence of the Bayesian approach is to provide a mathematical rule explaining how you change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. . . .

The Economist, September 30th 2000

Bayes Theorem I

Named after the English theologian and mathematician Thomas Bayes [1701–1761]

The theorem relies on the asymmetry of the definition of conditional probabilities:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \Rightarrow \mathrm{P}(A \cap B) = \mathrm{P}(B)\,\mathrm{P}(A|B) \qquad (1)$$

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \Rightarrow \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) \qquad (2)$$

for any two events A and B under regularity conditions, i.e. $\mathrm{P}(B) \neq 0$ in **??** and $\mathrm{P}(A) \neq 0$ in **??**.

Bayes Theorem II

Thus, from $\mathrm{P}(A|B)\,\mathrm{P}(B) = \mathrm{P}(B|A)\,\mathrm{P}(A)$ follows

Bayes Theorem

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)} \overset{\text{Law of tot. prob.}}{=} \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B|A)P(A) + \mathrm{P}(B|\bar{A})\,\mathrm{P}(\bar{A})}$$

More general, let $A_1, \ldots, A_n$ be *exclusive* and *exhaustive* events (ie they are a *partition* of the sample space), then

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}{\sum_{i=1}^{n} \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}$$

Bayes Theorem II

Thus, from $\mathrm{P}(A|B)\,\mathrm{P}(B) = \mathrm{P}(B|A)\,\mathrm{P}(A)$ follows

Bayes Theorem

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)} \overset{\text{Law of tot. prob.}}{=} \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B|A)P(A) + \mathrm{P}(B|\bar{A})\,\mathrm{P}(\bar{A})}$$

More general, let $A_1, \ldots, A_n$ be *exclusive* and *exhaustive* events (ie they are a *partition* of the sample space), then

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}{\sum_{i=1}^{n} \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}$$

Interpretation

$\mathrm{P}(A_i)$ prior probabilities

$\mathrm{P}(A_i|B)$ posterior probabilities

After observing B the prob. of A_i changes from $\mathrm{P}(A_i)$ to $\mathrm{P}(A_i|B)$.

Towards inference

A more general formulation of Bayes theorem is given by

$$f(X = x|Y = y) = \frac{f(Y = y|X = x)f(X = x)}{f(Y = y)}$$

where X and Y are random variables.

(Note: Switch of notation from P(.) to $f(.)$ to emphasise that we do not only relate to probabilities of events but to general probability functions of the random variables X and Y.)

Even more compact version

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}.$$

Bayesian Concepts

Also parameters are stochastic variables!

Bayesian Concepts

Also parameters are stochastic variables!

Example:

$$X \sim \text{Binom}(x; n, p)$$

Bayesian Concepts

Also parameters are stochastic variables!

Example:

$$X \sim \text{Binom}(x; n, p)$$

From basic course in statistics (classical/frequentist statistics):

- X is a stochastic variable with binomial distribution
- n is the numer of trials (known)
- p is a parameter, this is *unknown* but *fixed*

Bayesian Concepts

Also parameters are stochastic variables!

Example:

$$X \sim \text{Binom}(x; n, p)$$

From basic course in statistics (classical/frequentist statistics):

- X is a stochastic variable with binomial distribution
- n is the numer of trials (known)
- p is a parameter, this is *unknown* but *fixed*

In Bayesian statistics:

- p is a parameter, it is also a stochastic variable, it has a distribution $f(p)$
- The likelihood of X is seen as a conditional probability $P(X = x|p)$

Posterior distribution

Let:

- $X = x$ be the observed realization of a RV
- Assume $X \sim f(x|\theta)$ [Likelihood model]
- Assume $\theta \sim f(\theta$ [Prior Model]

The Bayes theorem allowes us to compute the posterior distribution

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}.$$

(For discrete parameter space the integral has to be replaced with a sum.)

The posterior distribution is the most important quantity in Bayesian inference. It contains all information about the unknown parameter θ after having observed the data $X = x$.

Posterior distribution (II)

Since the denominator in

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}$$

does not depend on θ, the density of the posterior distribution is proportional to

$$\underbrace{f(\theta|x)}_{\text{Posterior}} \propto \underbrace{f(x|\theta)}_{\text{Likelihood}} \times \underbrace{f(\theta)}_{\text{Prior}}$$

where $1/\int f(x|\theta)f(\theta)d\theta$ is the corresponding normalising constant to ensure $\int f(\theta|x)d\theta = 1$.

Binomial experiment

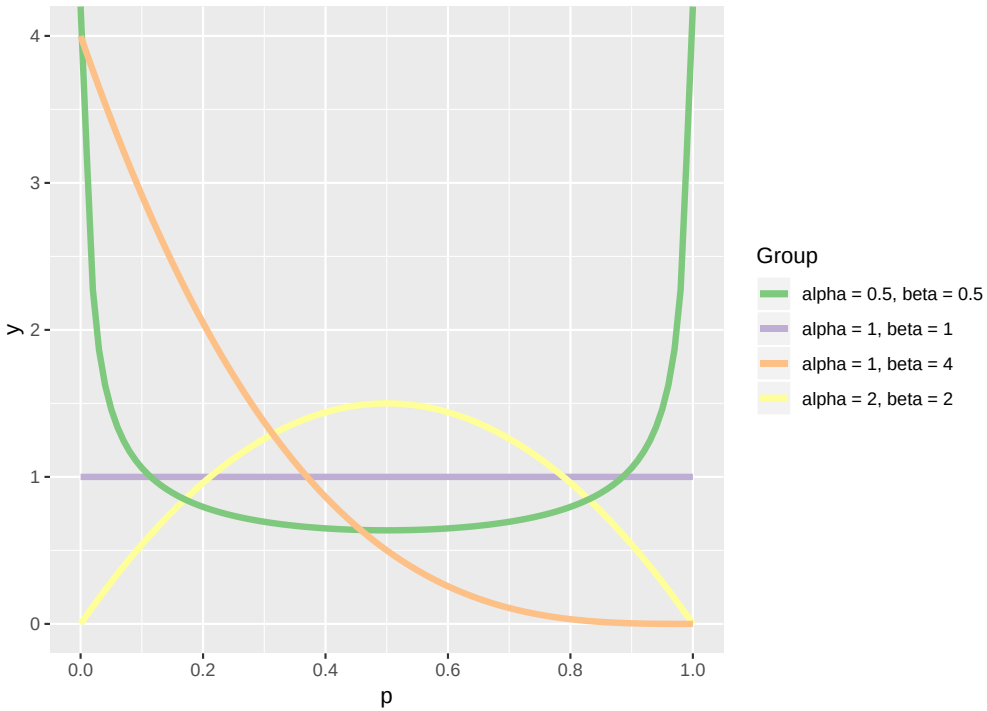
Let $X \sim \text{Bin}(n, p)$ with n known and unknown $p \in [0, 1]$.

Observe $x_1, \ldots, x_n \sim \text{Bin}(n, p)$ and assume iid.

Goal: estimate p given the data we have observed

Binomial experiment - Bayesian view

- Choose a prior for p.
- $p \sim \text{Be}(\alpha, \beta)$ is a common choice



Binomial experiment (2)

$$X \sim \text{Bin}(n, p),\ x = 0, 1, \ldots, n, \qquad p \sim \text{Be}(\alpha, \beta),\ 0 < p < 1$$

$$\Downarrow \qquad\qquad\qquad\qquad\qquad\qquad \Downarrow$$

Binomial experiment (2)

$$X \sim \text{Bin}(n, p),\ x = 0, 1, \ldots, n, \qquad p \sim \text{Be}(\alpha, \beta),\ 0 < p < 1$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\text{L}(p) \propto p^x (1-p)^{n-x} \qquad\qquad f(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

Binomial experiment (2)

$$X \sim \text{Bin}(n, p),\ x = 0, 1, \ldots, n, \qquad p \sim \text{Be}(\alpha, \beta),\ 0 < p < 1$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\text{L}(p) \propto p^x (1-p)^{n-x} \qquad\qquad f(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

Thus, the posterior distribution results as:

$$\begin{aligned} f(p|x) &\propto f(x|p) \times f(p) \\ &= p^x (1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \end{aligned}$$

Binomial experiment (2)

$$X \sim \text{Bin}(n, p),\ x = 0, 1, \ldots, n, \qquad p \sim \text{Be}(\alpha, \beta),\ 0 < p < 1$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\text{L}(p) \propto p^x (1-p)^{n-x} \qquad\qquad f(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

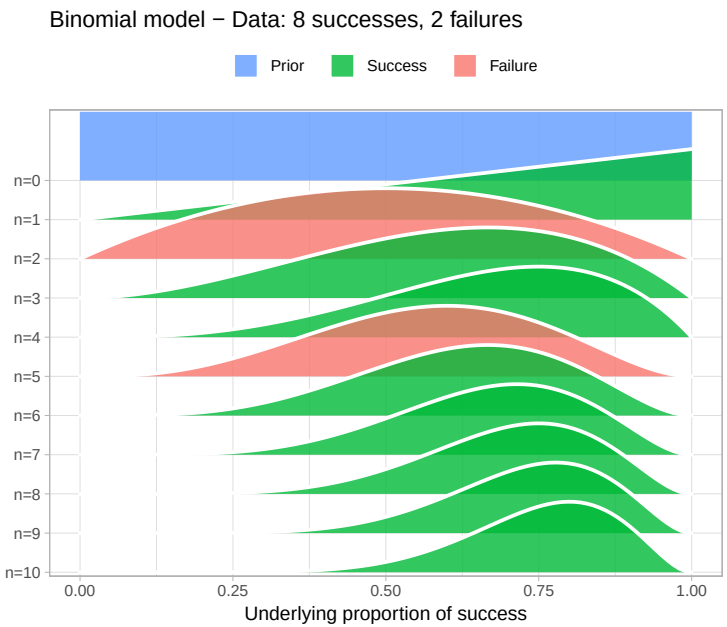
Thus, the posterior distribution results as:

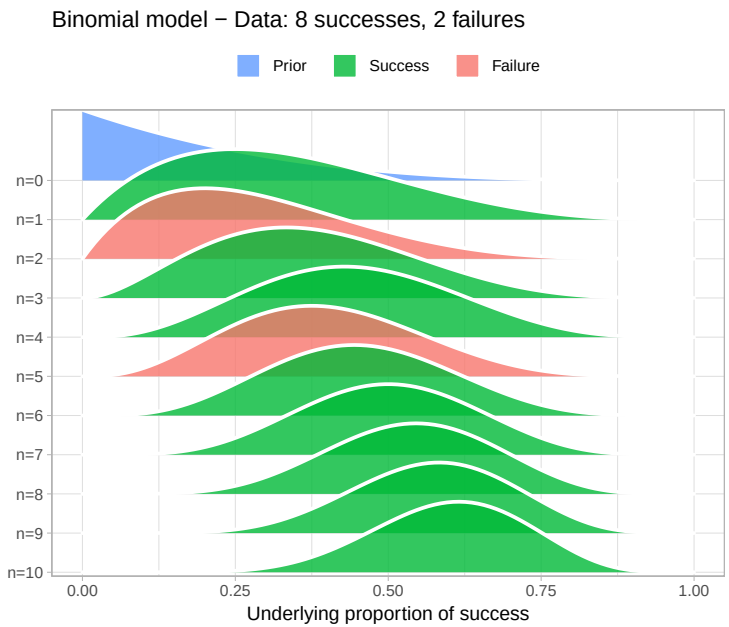
$$\begin{aligned} f(p|x) &\propto f(x|p) \times f(p) \\ &= p^x (1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \end{aligned}$$

This corresponds to the core of a beta distribution, so that

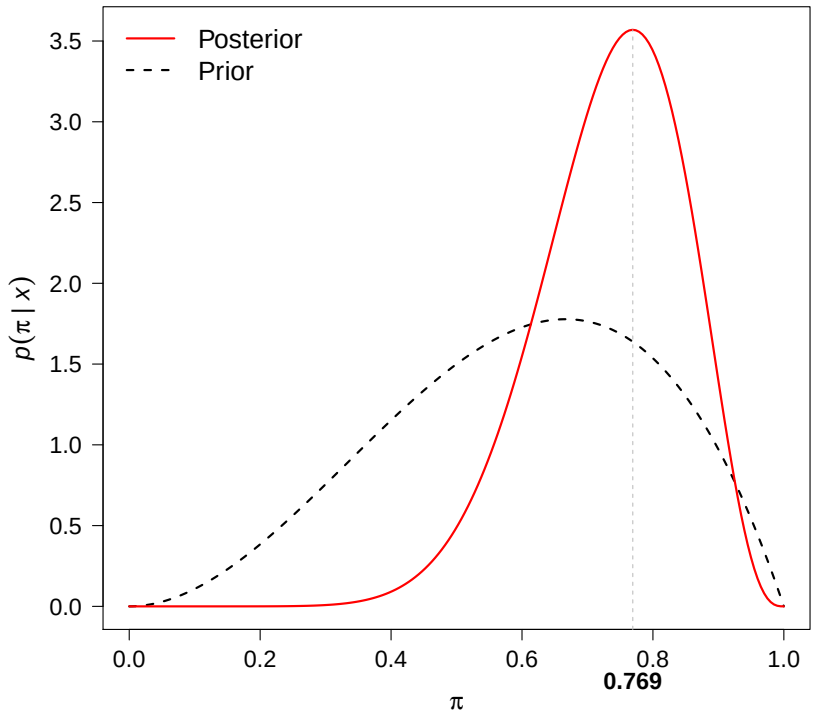
$$p|x \sim \text{Be}(\alpha + \underbrace{x}_{\text{successes}}, \beta + \underbrace{n-x}_{\text{failures}})$$

Binomial experiment: Simple example





Bayesian Inference



Posterior density of $p|x$ for a $\text{Be}(3, 2)$ prior and observation $x = 8$ in a binomial experiment with $n = 10$ trials.

Bayesian point estimates

Statistical inference about θ is based solely on the posterior distribution $f(\theta|x)$. Suitable point estimates are location parameters, such as:

- Posterior mean $\mathsf{E}(\theta|x)$:

$$\mathsf{E}(\theta|x) = \int \theta f(\theta|x) d\theta.$$

Bayesian point estimates

Statistical inference about θ is based solely on the posterior distribution $f(\theta|x)$. Suitable point estimates are location parameters, such as:

- Posterior mean $\mathrm{E}(\theta|x)$:

$$\mathrm{E}(\theta|x) = \int \theta f(\theta|x) d\theta.$$

- Posterior mode $\mathrm{Mod}(\theta|x)$:

$$\mathrm{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)$$

Bayesian point estimates

Statistical inference about θ is based solely on the posterior distribution $f(\theta|x)$. Suitable point estimates are location parameters, such as:

- Posterior mean $\mathrm{E}(\theta|x)$:

$$\mathrm{E}(\theta|x) = \int \theta f(\theta|x) d\theta.$$

- Posterior mode $\mathrm{Mod}(\theta|x)$:

$$\mathrm{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)$$

- Posterior median $\mathrm{Med}(\theta|x)$ is defined as the value a which satisfies

$$\int_{-\infty}^{a} f(\theta|x) d\theta = 0.5 \quad \text{and} \quad \int_{a}^{\infty} f(\theta|x) d\theta = 0.5$$

Credible interval

For fixed $\alpha \in (0, 1)$, a $(1 - \alpha)$ credible interval is defined through two real numbers t_l and t_u, so that

$$\int_{t_l}^{t_u} f(\theta|x)d\theta = 1 - \alpha.$$

The number $1 - \alpha$ is called the credible level of the credible interval $[t_l, t_u]$.

There are infinitely many $(1 - \alpha)$-credible intervals for fixed α. (At least if θ is continuous.)

Credible interval (II)

Equi-tailed credible interval

The same amount ($\alpha/2$) of probability mass is cut from the left and right tail of the posterior distribution, i.e. choose t_l as the $\alpha/2$-quantile and t_u as the $1-\alpha/2$-quantile.

Credible interval (II)

Equi-tailed credible interval

The same amount ($\alpha/2$) of probability mass is cut from the left and right tail of the posterior distribution, i.e. choose t_l as the $\alpha/2$-quantile and t_u as the $1-\alpha/2$-quantile.

Highest posterior density (HPD) intervals

Feature: The posterior density at any value of θ inside the credible interval must be larger than anywhere outside the credible interval. HPD-interval have the smallest width among all $(1-\alpha)$ credible intervals. For symmetric posterior distributions HPD intervals are also equi-tailed.

Binomial Esperiment - Confidence Interval

Let $X_1, \ldots, X_n \sim \text{Bin}(p, n)$ and indepedent.

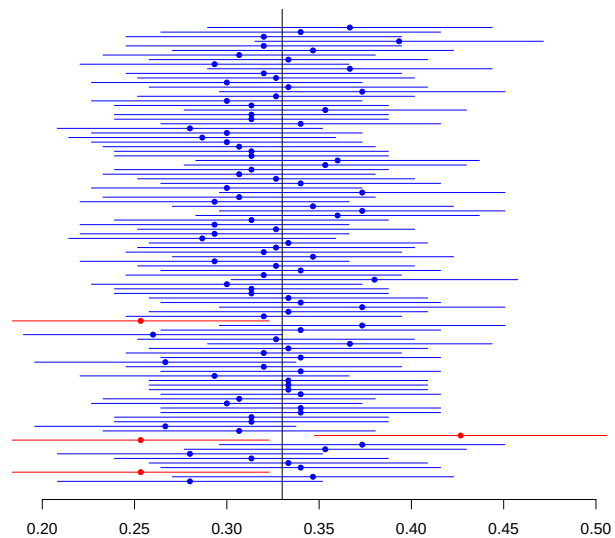
We have that for large n

$$\hat{p} = \frac{X}{n} \approx \mathcal{N}(p, \frac{p(1-p)}{n})$$

A confidence interval is then

$$\hat{p} \pm z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Interpretation: The interval as probability α of covering the true value of p



Bayesian Inference

All inference is based on the posterior distribution

$$f(p|x) \propto f(x|p)f(p)$$

- Point estimate: mean, mode, ...
- Interval estimate: choose t_l and t_u such that

$$P_{f(p|x)}(p \in [t_l, t_u]) = \alpha$$