Lecture 5: Review

What have we done until now?

- Simulation from discrete probability models
- Simulation from continuous probability models
  - Inversion Sampling and use known relationships between RV
  - Rejection Sampling
  - Monte Carlo Integration
  - Importance Sampling

Today

- Monte Carlo integration
- Importance sampling
- Bayesian statistics

Rejection sampling

- We want $x \sim f(x)$ (target density).
- We know how to generate realisations from a density $g(x)$
- We know a value $c > 1$, so that $\frac{f(x)}{g(x)} \leq c$ for all x where $f(x) > 0$.

Algorithm:

```
finished = 0
while (finished = 0)
    generate $x \sim g(x)$
    compute $\alpha = \frac{1}{c} \cdot \frac{f(x)}{g(x)}$
    generate $u \sim U[0, 1]$
    if $u \leq \alpha$ set finished = 1
return $x$
```

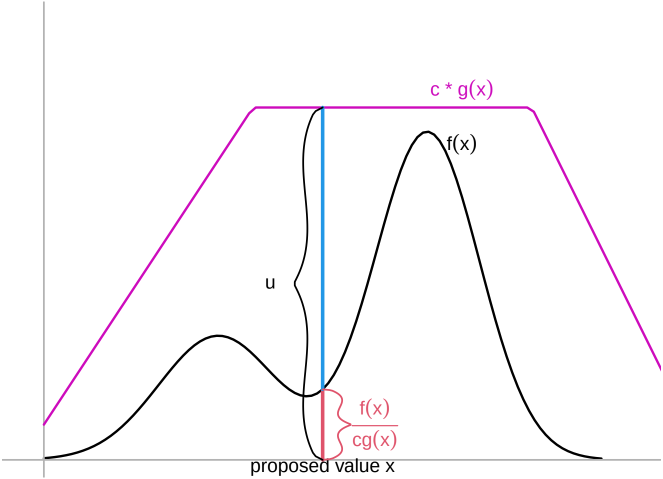
Rejection Sampling

- Overall acceptance probability: $1/c$
- Mean number of generate samples per accepted sample: c
- Don't need to know the normalizing constant
- Often not efficient in high dimensions

Difficulties:

- Find the constant $c \longrightarrow$ Sample importance resampling
- Find a good proposal distribution $\longrightarrow$ adaptive importance sampling

Rejection sampling



Monte Carlo integration

Assume we are interested in

$$\mu = \mathsf{E}[h(X)]; \; X \sim f(x)$$

If X is continuous and scalar we have

$$\mu = \mathsf{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) \, dx$$

Analytical solution is the best when possible!

Monte Carlo integration

Assumption

It is *easy* to generate independent samples $x_1, \ldots, x_N$ from a distribution $f(x)$ of interest.

A Monte Carlo estimate of

$$\mu = \mathrm{E}(h(x)) = \int h(x) f(x) dx$$

is then given by

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} h(x_i).$$

What is the mean and variance of this estimator?

Monte Carlo integration (II)

$\hat{\mu}$ is an unbiased estimate of μ

- $\mathrm{E}(\hat{\mu}) = \mu$
- $\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (h(x_i) - \hat{\mu})^2$
- Then the strong law of large numbers says:

$$\hat{\mathrm{E}}(h(x)) = \frac{1}{N} \sum_{i=1}^{N} h(x_i) \overset{a.s}{\to} \int h(x) f(x) dx = \mathrm{E}(h(x))$$

Note: Independent samples are not necessary for simulation-consistency.

However, the accuracy of a Monte Carlo estimate will be reduced if the samples are positively correlated.

Monte Carlo integration (III)

Monte carlo integration can be used for any function $h(\cdot)$

Examples

- Using $h(x) = x^2$ we obtain an estimate for $\mathrm{E}(x^2)$.
- An estimate for the variance follows as

$$\widehat{\mathrm{Var}}(x) = \hat{\mathrm{E}}(x^2) - \hat{\mathrm{E}}(x)^2$$

- Setting $h(x) = I(x \in A)$ we get:

$$E[h(x)] = E[I(x \in A)] = \mathrm{P}(x \in A)$$

Importance sampling

One of the principal reasons for wishing to sample from complicated probability distributions $f(z)$ is to be able to evaluate expectations with respect to some function $p(z)$:

$$\mathrm{E}(p) = \int p(z) f(z) dz$$

The technique of importance sampling provides a framework for approximating expectations directly but does not itself provide a mechanism for drawing samples from a distribution.

Importance sampling: Idea

[See notes (from today and from lecture 4)]

Importance sampling

Let $x_1, \ldots, n_N \sim g(x)$ then the importance sampling estimator of $\mu = \mathrm{E}_f(h(x))$ is given by

$$\hat{\mu}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{h(x_i) f(x_i)}{g(x_i)} = \frac{1}{N} \sum_{i=1}^{N} h(x_i) w(x_i)$$

wih

- We need $g(x) > 0$ where $h(x) f(x) > 0$
- The quantities $w(x_i) = \frac{f(x_i)}{g(x_i)}$ are called importance weights
- $\mathrm{E}(\hat{\mu}_{IS}) = \mu$
- $\mathrm{Var}(\hat{\mu}_{IS}) = \frac{1}{N} \mathrm{Var}_g[\frac{h(x) f(x)}{g(x)}]$

Importance sampling estimators

To compute the importance sampling estimator

$$\hat{\mu}_{IS} = \frac{1}{N}\sum_{i=1}^{N} h(x_i)w(x_i)$$

we need to know the normalizing constant of f and g.

When this is not possible an alternative is a "self-normalizing" or "reweighted" importance sampling estimator

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

where we need that

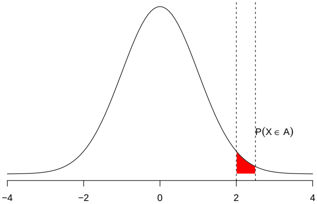
$$g(x) > 0 \text{ where } f(x) > 0$$

Importance sampling: Example

Assume we want to estimate

$$P(X \in [2, 2.5]) \text{ where } X \sim \mathcal{N}(0, 1)$$

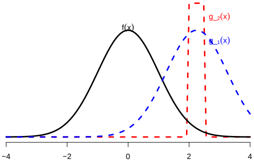
- Can use MC estimate → small efficiency
- Importance sampling can help "focus" the sampler in the correct area



`Show code`

Importance sampling: Example

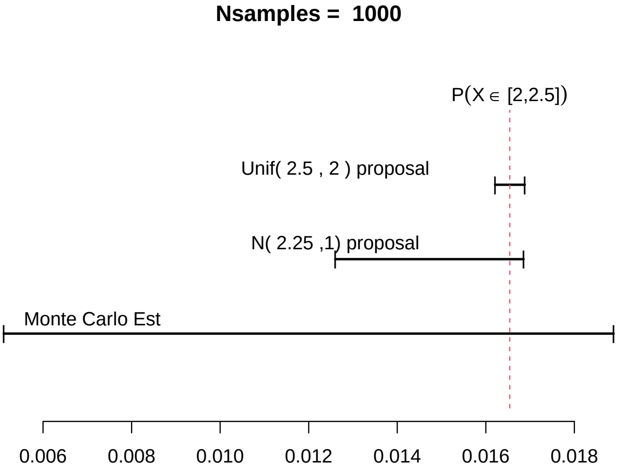
$\mu = P(X \in [2, 2.5]) = \int_{\mathcal{R}} I(x \in [2, 2.5]) f(x) dx$ with $f(x) = \mathcal{N}(0, 1)$
Three estimation schemes:



1. MC estimate
2. IS with proposal $g_1(x) = \mathcal{N}(2.25, 1)$
3. IS with proposal $g_2(x) = \mathcal{U}(2, 2.5)$

Note: in case 3) we cannot use the self-normalizing version of the IS algorithm

Importance sampling: Example



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

Importance sampling: Summary

As with rejection sampling, the success of importance sampling depends crucially on how well the proposal distribution $g(x)$ matches the target distribution $f(x)$.

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 (1) and $\mathrm{P}(A) \neq 0$ in (2).

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!

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 $\int f(\theta|x)d\theta = 1$.

Example: Binomial experiment

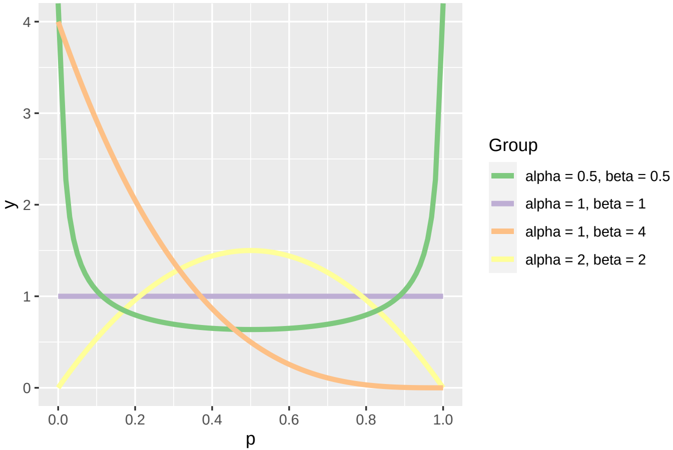
Let $X \sim \text{Bin}(n, p)$ with n known and unknown $p \in [0, 1]$.

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.
- $p \sim \text{Beta}(\alpha, \beta)$ is a common choice



Binomial experiment (2)

$$\begin{array}{ccc} X \sim \mathrm{Bin}(n,p),\ x=0,1,\ldots,n, & & p \sim \mathrm{Be}(\alpha,\beta),\ 0<p<1 \\ \Downarrow & & \Downarrow \\ \mathrm{L}(p) \propto p^{x}(1-p)^{n-x} & & f(p) \propto p^{\alpha-1}(1-p)^{\beta-1} \end{array}$$

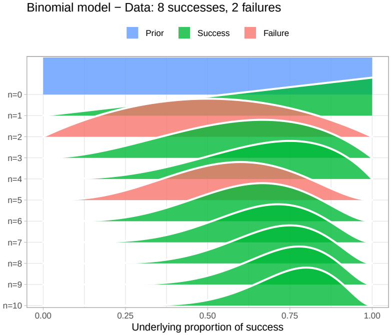
Thus, the posterior distribution results as:

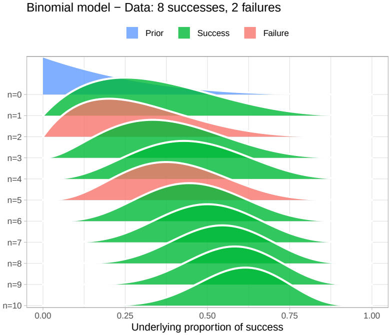
$$\begin{aligned} f(p|x) &\propto f(x|p) \times f(p) \\ &= p^{x}(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \end{aligned}$$

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

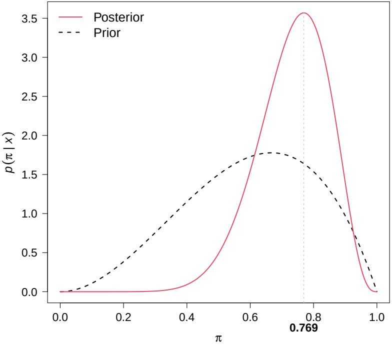
$$p|x \sim \mathrm{Be}(\alpha + \underbrace{x}_{\text{successes}}, \beta + \underbrace{n-x}_{\text{failures}})$$

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

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

- 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)

Equal-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 Experiment - Confidence Interval

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

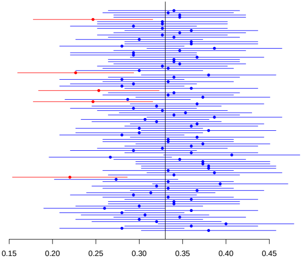
We have that for large n

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

A confidence interval is then

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

(Wrong) interpretation: The interval has probability $1-\alpha$ of covering the true value of p



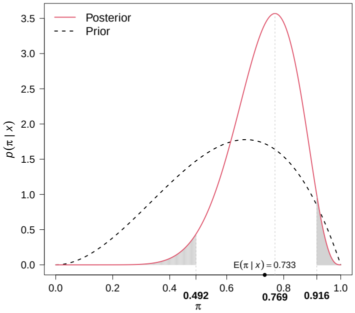
Bayesian Inference

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

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



Bayesian learning

An important feature of Bayesian inference is the consistent processing of sequentially arising data.

- Suppose new independent data x_2 from a $\text{Bin}(n, p)$ arrive.
- The posterior distribution from the original observation (with x now called x_1) becomes the prior for x_2:

$$\begin{aligned} f(p|x_1, x_2) &\propto f(x_2|p, x_1) \times f(p|x_1) \\ &\propto f(x_2|p) \times f(p|x_1) \end{aligned}$$

Using $f(p|x_1) \propto f(x_1|p) \times f(p)$ an alternative formula is

$$\begin{aligned} f(p|x_1, x_2) &\propto f(x_2|p) \times f(x_1|p) \times f(p) \\ &= f(x_1, x_2|p) \times f(p) \end{aligned}$$

Thus, $f(p|x_1, x_2)$ is the same whether or not the data are processed sequentially.

Choice of the prior distribution

Prior distributions incorporate prior beliefs in the Bayesian analysis.

A pragmatic approach is to choose a conjugate prior distribution.

Conjugate prior distribution

Let $\mathrm{L}_x(\theta) = \mathrm{p}(x|\theta)$ denote a likelihood function based on the observation $X = x$. A class $\mathcal{G}$ of distributions is called conjugate with respect to $\mathrm{L}_x(\theta)$ if the posterior distribution $\mathrm{p}(\theta|x)$ is in $\mathcal{G}$ for all x whenever the prior distribution $\mathrm{p}(\theta)$ is in $\mathcal{G}$.

Example

Binomial experiment: Let $X|p \sim \mathrm{Bin}(n, p)$. The family of beta distributions, $p \sim \mathrm{Be}(\alpha, \beta)$, is conjugate with respect to $\mathrm{L}_x(p)$, since the posterior distribution is again a beta distribution:

$p|x \sim \mathrm{Be}(\alpha + x, \beta + n - x)$

List of conjugate prior distributions

Sequential processing:

- Sufficient to study conjugacy for one member of a random sample $X_1, \ldots, X_n$.
- The posterior after observing the first observation is of the same type as the prior and serves as new prior distribution for the next observation.
- Sequentially processing the data, only the parameters will change and not the type of prior.

List of conjugate prior distributions

Likelihood	**Conjugate prior**	**Posterior distribution**
$X\|p \sim \text{Bin}(n, p)$	$p \sim \text{Be}(\alpha, \beta)$	$p\|x \sim \text{Be}(\alpha + x, \beta + n - x)$
$X\|p \sim \text{Geom}(p)$	$p \sim \text{Be}(\alpha, \beta)$	$p\|x \sim \text{Be}(\alpha + 1, \beta + x - 1)$
$X\|\lambda \sim \text{Po}(e \cdot \lambda)$	$\lambda \sim \text{G}(\alpha, \beta)$	$\lambda\|x \sim \text{G}(\alpha + x, \beta + e)$
$X\|\lambda \sim \text{Exp}(\lambda)$	$\lambda \sim \text{G}(\alpha, \beta)$	$\lambda\|x \sim \text{G}(\alpha + 1, \beta + x)$
$X\|\mu \sim \mathcal{N}(\mu, \sigma^2_\star)$	$\mu \sim \mathcal{N}(\nu, \tau^2)$	$\mu\|x \sim \mathcal{N}\left[(A)^{-1}\left(\frac{x}{\sigma^2} + \frac{\nu}{\tau^2}\right), (A)^{-1}\right]$
$X\|\sigma^2 \sim \mathcal{N}(\mu_\star, \sigma^2)$	$\sigma^2 \sim \text{IG}(\alpha, \beta)$	$\sigma^2\|x \sim \text{IG}(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2)$

$_\star$: known.

$A = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$

Improper prior distributions

Maybe you feel uncomfortable putting a prior on an unknown parameter. If you use a normal prior you can use a very large variance. In the limit this leads to an improper prior distribution.

Improper prior distribution

For example, let $\mu \sim \mathcal{N}(\mu, \infty)$, i.e. $p(\mu) \propto \text{const.} > 0$.

$$\int p(\mu) d\mu \approx \infty$$

Priors such as $p(\mu) = \text{const.}, p(\sigma) = 1/\sigma$ are improper, because they do not integrate to 1.

Improper prior distributions (II)

In most cases, improper priors can be used in Bayesian analyses without major problems. However, things to watch out for are:

- In a few models, the use of improper priors can result in improper posteriors.
- Use of improper priors makes model selection difficult.

Uninformative priors

Though conjugate priors are computationally nice, priors might be preferred which do not strongly influence the posterior distribution. Such a prior is called an uninformative prior.

- The historical approach, followed by Laplace and Bayes, was to assign flat priors.
- This prior seems reasonably uninformative. We do not know where the actual value lies in the parameter space, so we might as well consider all values equi-probable.
- However, this prior is not invariant to one-to-one transformations.

HAROLD JEFFREYS' PRIOR

Definition

Let X denote a random variable with likelihood function $p(x|\theta)$ where θ is an unknown scalar parameter. Jeffreys' prior or Jeffreys' rule is defined as

$$\mathrm{p}(\theta) \propto \sqrt{J(\theta)},$$

where $J(\theta)$ is the expected Fisher information of θ.

Jeffreys' prior has certain desired properties, e.g. invariance property.

Jeffreys' prior for the geometric distribution

The geometric distribution models the number X of Bernoulli trials needed to get the first success. Let $X|\pi \sim \text{Geom}(\pi)$, i.e.

$$f(x|\pi) = \pi \cdot (1-\pi)^{x-1}, \quad \text{so}l_x(\pi) \quad = \log(\pi) + (x-1)\log(1-\pi).$$

Jeffreys' prior for the geometric distribution (II)

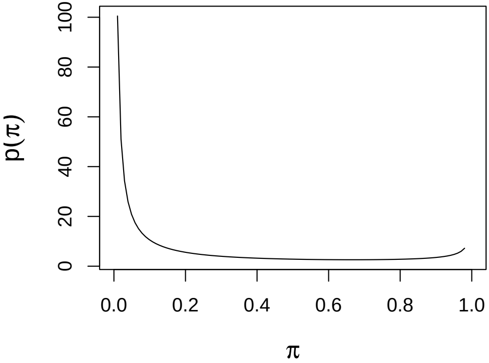
$$
\begin{aligned}
l_x(\pi) &= \log(\pi) + (x-1)\log(1-\pi) \\
l_x'(\pi) &= \frac{1}{\pi} - \frac{x-1}{1-\pi} \\
l_x''(\pi) &= -\frac{1}{\pi^2} - \frac{x-1}{(1-\pi)^2}
\end{aligned}
$$

$$
\begin{aligned}
J(\pi) &= -\,\mathsf{E}\left(-\frac{1}{\pi^2} - \frac{x-1}{(1-\pi)^2}\right) \\
&= \frac{1}{\pi^2} + \frac{\frac{1}{\pi}-1}{(1-\pi)^2} \\
&= \frac{1}{\pi^2} + \frac{1-\pi}{\pi(1-\pi)^2} \\
&= \pi^{-2}(1-\pi)^{-1}
\end{aligned}
$$

Jeffreys' prior for the geometric distribution (III)

Jeffreys' prior results as: $\mathrm{p}(\pi) \propto \sqrt{J(\pi)} = \pi^{-1}(1-\pi)^{-1/2}$

(can be seen as "$\mathrm{Be}(0, 0.5)$")



$\Rightarrow$ Small values are favoured.