Choosing Prior Distributions

Ben Goodrich

February 14, 2019

Obligatory Disclosure

- Ben is an employee of Columbia University, which has received several research grants to develop Stan
- Ben is also a manager of GG Statistics LLC, which uses Stan for business purposes
- According to Columbia University policy, any such employee who has any equity stake in, a title (such as officer or director) with, or is expected to earn at least $\$5,000.00$ per year from a private company is required to disclose these facts in presentations

Analytical Posterior PDF

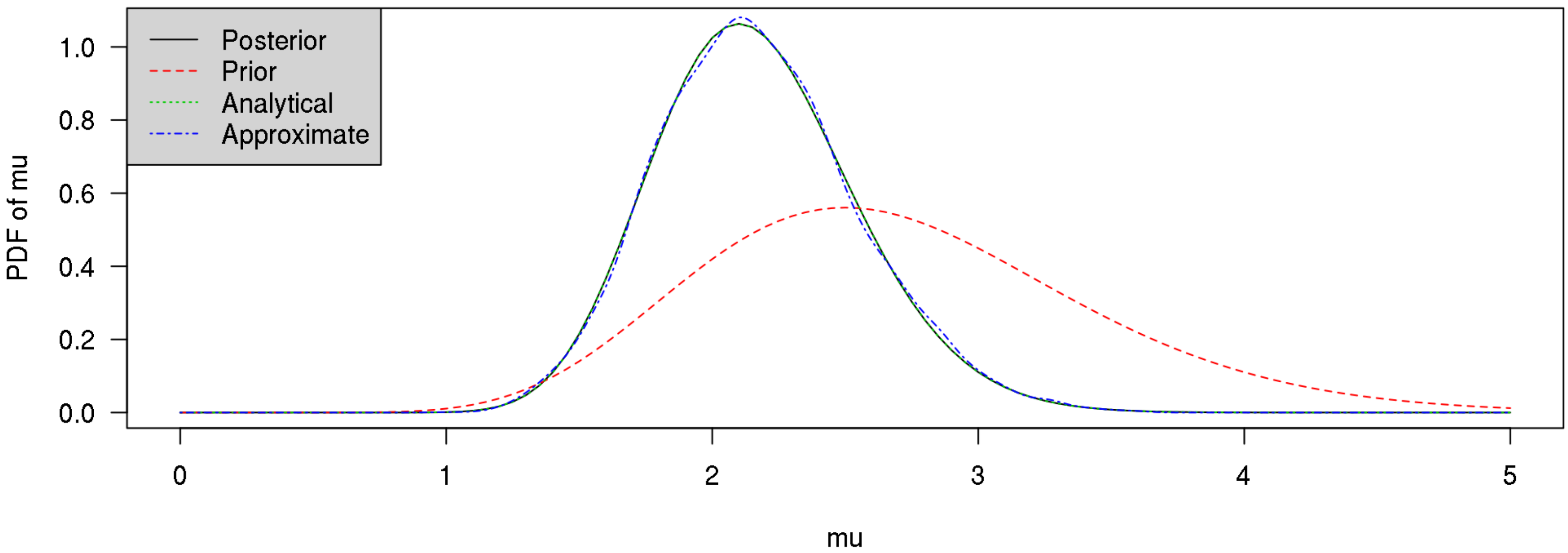
- Gamma prior PDF is again $f\left(\mu \middle| a, b\right) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}$
- Poisson PMF for N observations is again $f\left(y_1, \ldots, y_n \middle| \mu\right) = e^{-N\mu} \mu^{\sum_{n=1}^{N} y_n} \prod_{n=1}^{N} \frac{1}{y_n!}$
- Posterior PDF, $f\left(\mu \middle| a, b, y_1, \ldots, y_n\right)$, is proportional to their product:

$$\mu^{a-1} e^{-b\mu} \mu^{\sum_{n=1}^{N} y_n} e^{-N\mu} = \mu^{a-1+\sum_{n=1}^{N} y_n} e^{-(b+N)\mu} = \mu^{a^*-1} e^{-b^*\mu},$$

  where $a^* = a + \sum_{n=1}^{N} y_n$ and $b^* = b + N$
- Ergo, the posterior has a Gamma kernel and the normalizing constant is $\frac{(b^*)^{a^*}}{\Gamma(a^*)}$

Posterior vs. Prior PDF

```
curve(kernel(mu, M, Emu, y) / denom, from = 0, to = 5, ylab = "PDF of mu", xname = "mu")
curve(dgamma(mu, a, b), lty = 2, col = 2, add = TRUE, xname = "mu")
curve(dgamma(mu, a + sum(y), b + length(y)), lty = 3, col = 3, add = TRUE, xname = "mu")
lines(density(post, from = 0, to = 5), lty = 4, col = 4)
legend("topleft", legend = c("Posterior", "Prior", "Analytical", "Approximate"),
       lty = 1:4, col = 1:4, bg = "lightgrey")
```

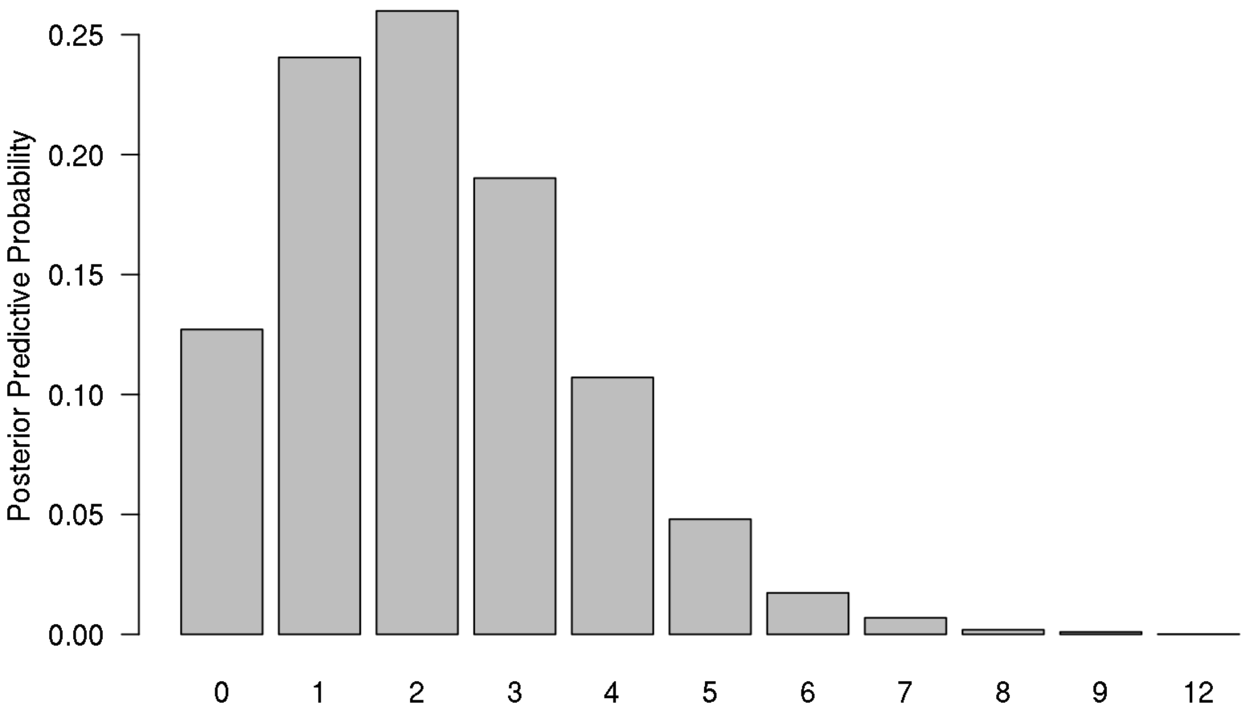


Posterior Predictive Distribution

- What do you believe a FUTURE outcome will be?
- Its PDF is $f\left(y \mid a, b, y_1, \ldots, y_n\right)=\int_0^{\infty} f\left(y, \mu \mid a, b, y_1, \ldots, y_n\right) d\mu =$ $\int_0^{\infty} f\left(y \mid \mu\right) f\left(\mu \mid a, b, y_1, \ldots, y_n\right) d\mu=\int_0^{\infty} \frac{e^{-\mu} \mu^y}{y!} \frac{\left(b^*\right)^{a^*}}{\Gamma\left(a^*\right)} \mu^{a^*-1} e^{-b^* \mu} d\mu=$ $\frac{\left(b^*\right)^{a^*}}{y!\Gamma\left(a^*\right)} \int_0^{\infty} \mu^{a^*+y-1} e^{-\left(b^*+1\right) \mu} d\mu=\frac{\left(b^*\right)^{a^*}}{y!\Gamma\left(a^*\right)} \frac{\Gamma\left(a^*+y\right)}{\left(b^*+1\right)^{a^*+y}}=\frac{\Gamma\left(a^*+y\right)}{y!\Gamma\left(a^*\right)}\left(\frac{b^*}{b^*+1}\right)^{a^*} \frac{1}{\left(b^*+1\right)^y}$
- This is ONE way to write the PMF for the negative binomial distribution over the non-negative integers, which has expectation $\frac{a^*}{b^*}$ but variance $\frac{a^*}{b^*}+\frac{a^*}{b^* b^*}$ that is larger than the expectation because you are not certain about μ

Drawing from a Posterior Predictive Distribution

```r
y_tilde <- rpois(S, post) # R functions that generate random numbers start with r
barplot(prop.table(table(y_tilde)), ylab = "Posterior Predictive Probability")
```



Four Ways to Execute Bayes Rule

1. Draw from the prior predictive distribution and keep realizations of the parameters iff the realization of the outcome matches the observed data
   - Very intuitive what is happening but is only possible with discrete outcomes and only feasible with few observations and parameters

2. Numerically integrate the numerator of Bayes Rule over the parameter(s)
   - Follows directly from Bayes Rule but is only feasible when there are few parameters and can be inaccurate even with only one parameter

3. Analytically integrate the numerator of Bayes Rule over the parameter(s)
   - Makes incremental Bayesian learning obvious but is only possible in simple models when the distribution of the outcome is in the exponential family

4. Use Stan to perform MCMC to sample from the posterior distribution
   - Works for any posterior PDF that is differentiable with respect to the parameters but can take a long time

Quantity of Interest for Bayesians & Frequentists

- Bayesians are ultimately interested in expectations of the form $\mathbb{E}_{\boldsymbol{\theta}|y_1\dots y_N} g(\boldsymbol{\theta}) = \int \dots \int g(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\, y_1 \dots y_N)\, d\theta_1 \dots d\theta_K$ where $g(\boldsymbol{\theta})$ is some function of the unknown parameters, such as utility for an action, and $f(\boldsymbol{\theta}|\, y_1 \dots y_N)$ is a posterior PDF for unknown parameters given $y_1 \dots y_N$
- Frequentists are ultimately interested in expectations of the form $\mathbb{E}_{Y|\boldsymbol{\theta}} h(y_1 \dots y_N) = \int \dots \int h(y_1 \dots y_N) f(y_1 \dots y_N |\, \boldsymbol{\theta})\, dy_1 \dots dy_N$ where $h(y_1 \dots y_N)$ is some function of the data, such as a point estimator of $\boldsymbol{\theta}$ and $f(y_1 \dots y_N |\, \boldsymbol{\theta})$ is a PDF for the data-generating process given $\boldsymbol{\theta}$
- If $h(y_1 \dots y_N |\, \boldsymbol{\theta}) = \mathbb{I}\{\underline{\theta}(y_1 \dots y_N) < \theta < \overline{\theta}(y_1 \dots y_N)\}$, what estimators $\underline{\theta}(y_1 \dots y_N)$ and $\overline{\theta}(y_1 \dots y_N)$ imply $\mathbb{E}_{Y|\boldsymbol{\theta}} h(y_1 \dots y_N) = 0.95$?
- If you can derive such functions, $\left[\underline{\theta}(y_1 \dots y_N), \overline{\theta}(y_1 \dots y_N)\right]$ is a 95% confidence interval estimator of the point θ

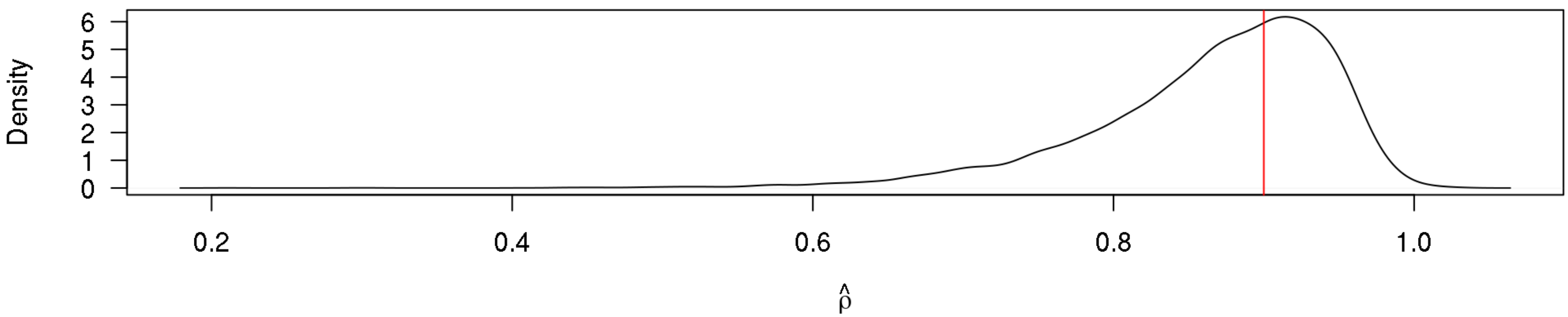
Frequentist Principles (Algorithm 1.4 in Lancaster)

```
functions { /* saved as AR1_rng.stan in R's working directory */
  vector AR1_rng(int S, int T, real mu, real rho, real sigma) {
    vector[S] rho_hat;  // holds OLS estimates of rho
    real alpha = mu * (1 - rho); int Tm1 = T - 1;
    if (sigma <= 0) reject("sigma must be positive");
    if (rho < -1 || rho > 1) reject("rho must be between -1 and 1");
    for (s in 1:S) { // repeatedly simulate data under an AR1 process ...
      vector[T] Y; Y[1] = 0;                                               // outcome at time 1
      for (t in 2:T) Y[t] = alpha + rho * Y[t - 1] + normal_rng(0, sigma);// outcome at time t
      { // ... and apply some function to that simulated data
        vector[Tm1] y_lag  = Y[1:Tm1]; vector[Tm1] y_temp = Y[2:T];
        rho_hat[s] = sum(y_temp .* y_lag) / sum(square(y_lag)); // .* multiplies elementwise
      }
    }
    return sort_asc(rho_hat);
  }
}
```

```
rstan::expose_stan_functions("AR1_rng.stan")
```

Sampling Distribution of the OLS Estimator of ρ

```
rho_hat <- AR1_rng(S = 10000L, T = 51, mu = 0, rho = 0.9, sigma = 1)
plot(density(rho_hat), main = "", xlab = expression(hat(rho))); abline(v = 0.9, col = 2)
```



The OLS estimator of ρ is biased (downward) because

$$\rho \neq \mathbb{E}_{Y|\rho}\left[\hat{\rho}\right] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \prod_{t=2}^{T} f\left(y_t \,|\, y_{t-1}, \mu, \rho, \sigma\right) dy_1 \dots dy_T$$

Important Points from Lancaster Chapter 1

- Statistics vs. Econometrics really is No Model vs. Generative Model
- Choosing YOUR prior is not fundamentally different from choosing YOUR likelihood (but it can be on behalf of someone else)
- Writing the normal distribution with μ and the "precision" $\tau = \frac{1}{\sigma^2}$
- Simulated data is more useful than wild data
- Likelihood Principle: "the data that might have been seen but were not are irrelevant!" and "Professional opinion is divided on whether inference should adhere to the likelihood principle."
- Bayesian inference does not require that the data be a "sample" from a well-defined population
- Identification: "A value θ_a of a parameter is identified if there is no other value θ_b such that $f\left(y \mid \theta_a\right) = f\left(y \mid \theta_b\right) \forall y \in \Omega$."
- The chapter appendix with several important probability distributions

Principles to Choose Priors (and likelihoods) with

1. Do not use improper priors
2. Subjective
3. Entropy Maximization
4. Invariance to reparameterization (particularly scaling)
5. “Objective” (actually also subjective, but different from 2)
6. Penalized Complexity (PC) (which we will cover when we get to hierarchical models)

Do Not Use Improper Priors

- Improper priors are those that do not have a PDF that integrates to 1
- Thus, you cannot draw from such priors or their prior predictive distributions
- In some situations, using an improper prior implies that the posterior distribution is improper and thus USELESS for Bayesian inference
- In other situations, an improper prior yields a proper posterior distribution but you have to work it out on a case-by-case basis
- Proper priors (that integrate to 1) ALWAYS yield proper posteriors
- Even if an improper prior yields a proper posterior distribution, the improper prior prelcudes model comparison via Bayes Factors
- Improper priors can also make things computationally problematic, so they are discouraged for people who use Stan

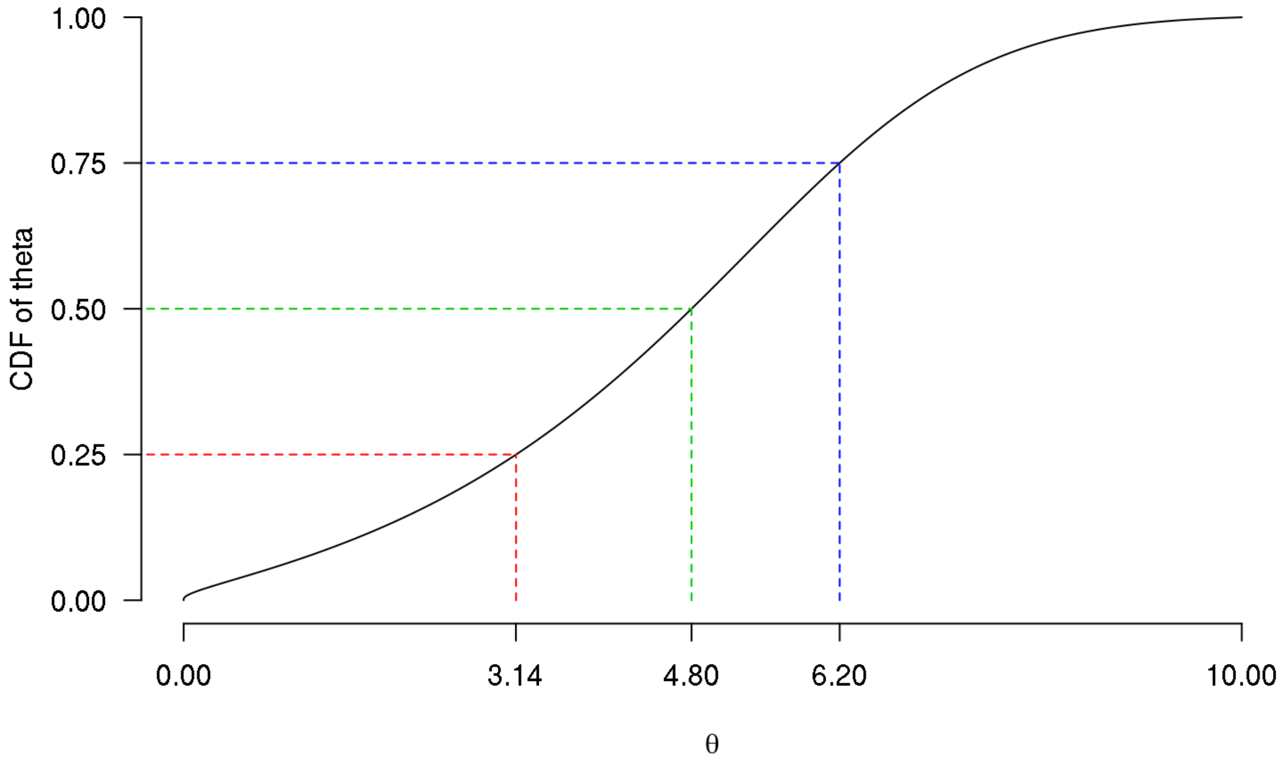
Subjective Priors

- Choose priors to reflect your (or your audience's) beliefs about the parameters
- This can include eliciting prior information from "experts"
- http://metalogdistributions.com/publications.html and JQPD.stan

Parameter Space	Required Inputs (besides $p \sim$ Uniform)	Function Name
$\Theta = \mathbb{R}$	$\alpha_j = \Pr\left(\theta \leq x_j\right)$ for $j = 1, 2, 3, 4$	`qnormal_icdf`
$\Theta = (l, u)$	l, u, $\alpha = \Pr\left(\theta \leq x_\alpha\right)$, $0.5 = \Pr\left(\theta \leq x_{0.5}\right)$, $1 - \alpha = \Pr\left(\theta \leq x_{1-\alpha}\right)$	`JQPDB_icdf`
$\Theta = (l, \infty)$	l, moments?, $\alpha = \Pr\left(\theta \leq x_\alpha\right)$, $0.5 = \Pr\left(\theta \leq x_{0.5}\right)$, $1 - \alpha = \Pr\left(\theta \leq x_{1-\alpha}\right)$	`JQPDS_icdf` or `JQPDS2_icdf`

Using Quantile Parameterized Distributions

- Alexa, show me a prior distribution over $\Theta = (0, 10)$ with a first quartile of π, a median of 4.8, and a third quartile of 6.2



Entropy Maximization

- One way of choosing a distribution: Choose $f(\theta|\cdot)$ to maximize $\mathbb{E}[-\ln f(\theta|\cdot)]$ subject to the restrictions that $\int_\Theta f(\theta|\cdot)\, d\theta = 1$ and $\int_\Theta g_j(\theta) f(\theta|\cdot)\, d\theta = m_j$ for one or more known values of m_j that correspond to the expectation of $g_j(\theta)$
- In the discrete case, a uniform distribution reaches the entropy upper bound
- By analogy, the maximum entropy distribution is the probability distribution "closest" to the uniform distribution while satisfying the constraints
- In other words, the maximum entropy distribution conveys the least amount of extra information about θ beyond the information that $\mathbb{E}g_j(\theta) = m_j$
- This process can be used to choose priors and / or likelihoods

Important Maximimum Entropy Distributions

- If Θ is some convex set, the maximum entropy distribution is the uniform distribution over Θ. For example, if $\Theta = [0,1]$, it is the standard uniform distribution with PDF $f\left(\theta \middle| a=0, b=1\right) = 1$
- If $\Theta = \mathbb{R}$, $m_1 = \mu$, and $m_2 = \sigma^2$, then the maximum entropy distribution is the normal distribution. This extends to bivariate and multivariate distributions if you have given covariances.
- If $\Theta = \mathbb{R}_+$ and $m_1 = \mu$, then the maximum entropy distribution is the exponential distribution with expectation $\mu = \frac{1}{\lambda}$. You can utilize the fact that the median is $F^{-1}\left(0.5\right) = \mu \ln 2$ to go from the median to μ.
- The binomial and Poisson distributions are maximum entropy distributions given μ for their respective Ω
- Additional examples (often with weird constraints) are given at the bottom of https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution

Invariance to Reparameterization

- A Jeffreys prior is proportional to the square root of the Fisher information
- The Fisher information is defined as $I(\theta) =$

$$-\mathbb{E}\left[\frac{\partial^2 \ell(\theta; y_1 \dots y_N)}{\partial\theta\partial\theta}\right] = -\int_\Omega \dots \int_\Omega \frac{\partial^2 \ell(\theta; y_1 \dots y_N)}{\partial\theta\partial\theta} f(y_1 \dots y_N | \theta)\, dy_1 \dots dy_N$$

  where $\ell(\theta; y_1 \dots y_N)$ is the log-likelihod of the sample of size N
- Jaynes argued that the Jeffreys prior really only makes sense for a scale parameter and in that case $f(\theta) \propto \frac{1}{\theta} = \sqrt{I(\theta)}$, which is improper
- The Jeffreys prior on a scale parameter is the non-informative prior that conveys the information that the units of θ convey no substantive information about its value, i.e. the Jeffreys prior is the same whether θ is in pounds or kilograms

“Objective” Priors

- “Objective” priors are not actually objective and they all convey some information that you choose to prioritize
- Reference priors choose $f\left(\theta\right| \cdot)$ such that the EXPECTED amount of information in the posterior that is contributed by the prior is minimal
- Reference priors do not always exist
- Reference priors can be very odd
- Reference priors often are the same as Jeffreys prior

Three “Uninformative” Beta Priors

- The beta distribution is the maximum entropy distribution for a given $\mathbb{E}\ln\theta$ and $\mathbb{E}\ln(1-\theta)$. If your beliefs are such that $\mathbb{E}\ln\theta = \mathbb{E}\ln(1-\theta)$, then $a = 1 = b$ and the beta distribution simplifies to the uniform on $\Theta = [0, 1]$
- But if the likelihood is binomial, then the posterior is beta with $a^* = a + y$ and $b^* = b + N - y$, so the uniform prior can be seen as adding one success and one failure to the likelihood. This denies that $\theta = 0$ and that $\theta = 1$
- Haldane thus argued the least informative beta prior was the limit as $a \downarrow 0$ and $b \downarrow 0$ at the same rate, which is a uniform prior on the log-odds $\eta = \frac{\theta}{1-\theta}$
- Jeffreys argued a reasonable way to construct a prior would convey the same amount of information about θ as η, leading to a beta prior with $a = 0.5 = b$