04 - Binomial model 01

Bayesian Statistics

Spring 2022-2023

Josep Fortiana

Matemàtiques - Informàtica UB

Monday, March 06, 2023

04 - Binomial model 01

Estimating a probability

Which is the least informative prior?

04 - Binomial model 01

Estimating a probability

Which is the least informative prior?

Bayesian Bernoulli model

Sample: $X = (X_1, \dots, X_n)$ iid $\sim \operatorname{Ber}(\theta)$.

Estimate the probability $\theta \in \Theta = (0, 1)$.

Prior distribution for θ: if no previous information, assume $\operatorname{Unif}(0, 1)$:

$$p(\theta) = 1, \quad 0 < \theta < 1.$$

Non-Informative Prior (NIP).

A family of prior distributions

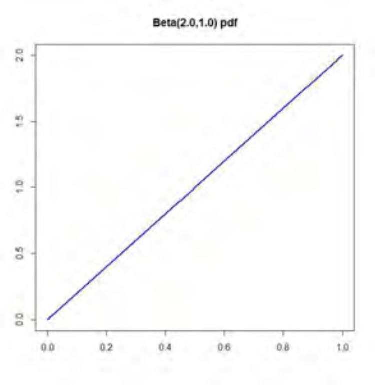
More generally: prior pdf of θ is $\mathrm{Beta}(\alpha, \beta)$:

$$p(t; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} t^{\alpha-1} (1-t)^{\beta-1}, \quad 0 < t < 1,$$

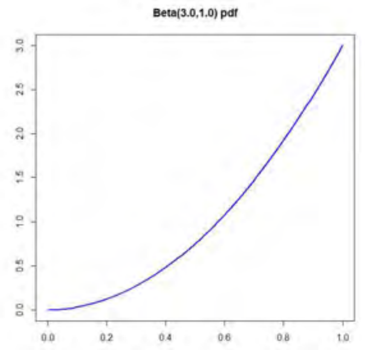
where $B(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$, is the Beta function.

In particular, $\mathrm{Beta}(1, 1) = \mathrm{Unif}(0, 1)$.

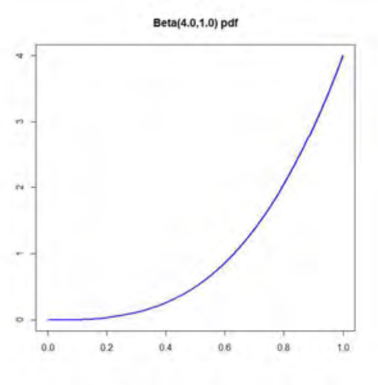
Examples of Beta pdf's with $\alpha, \beta \geq 1$



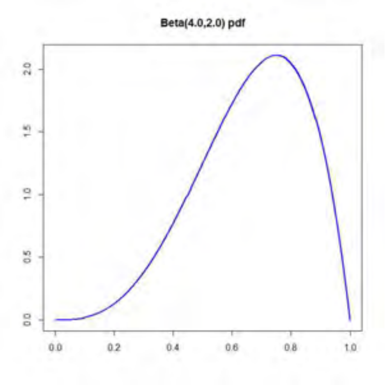
Examples of Beta pdf's with $\alpha, \beta \geq 1$



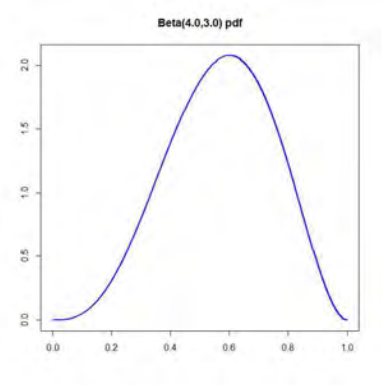
Examples of Beta pdf's with $\alpha, \beta \geq 1$



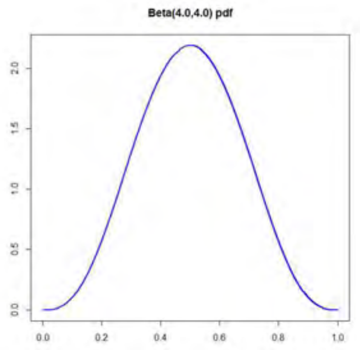
Examples of Beta pdf's with $\alpha, \beta \geq 1$



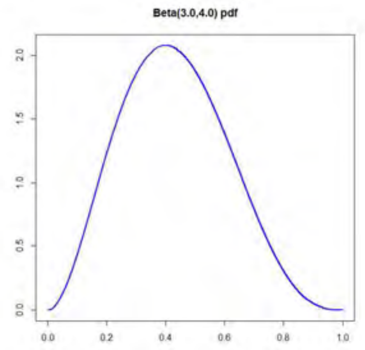
Examples of Beta pdf's with $\alpha, \beta \geq 1$



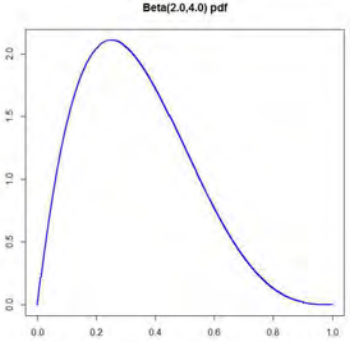
Examples of Beta pdf's with $\alpha, \beta \geq 1$



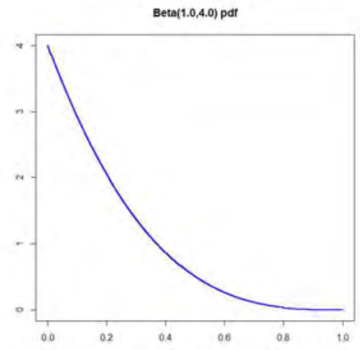
Examples of Beta pdf's with $\alpha, \beta \geq 1$



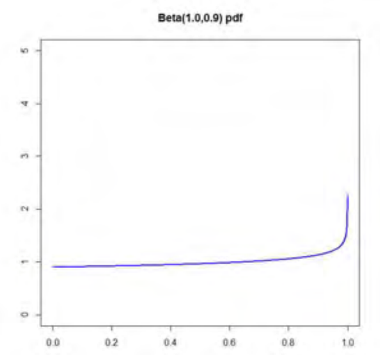
Examples of Beta pdf's with $\alpha, \beta \geq 1$



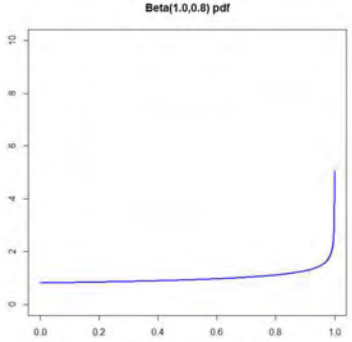
Examples of Beta pdf's with $\alpha, \beta \geq 1$



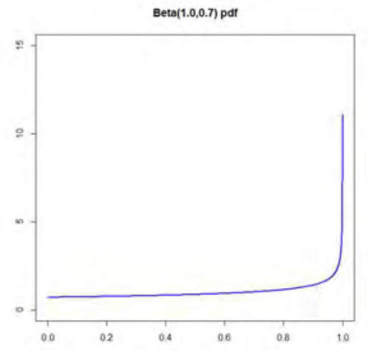
Examples of Beta pdf's with $\alpha, \beta < 1$



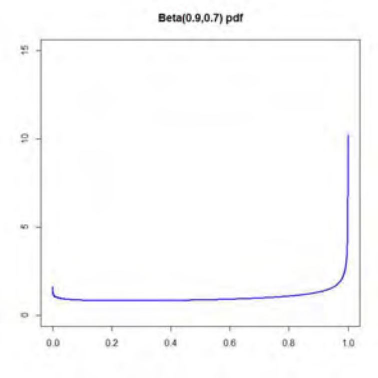
Examples of Beta pdf's with $\alpha, \beta < 1$



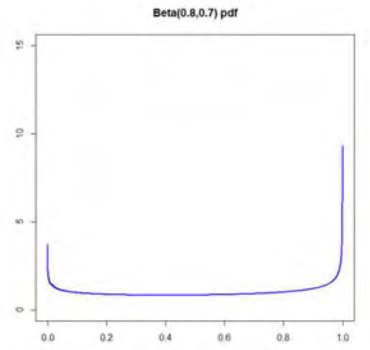
Examples of Beta pdf's with $\alpha, \beta < 1$



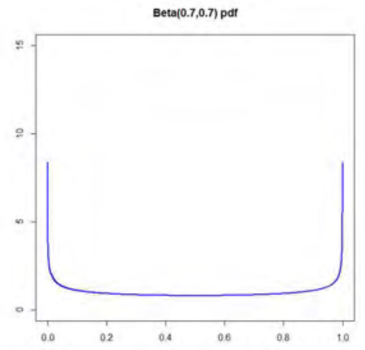
Examples of Beta pdf's with $\alpha, \beta < 1$



Examples of Beta pdf's with $\alpha, \beta < 1$



Examples of Beta pdf's with $\alpha, \beta < 1$



Likelihood

We observe n values $X_i = x_i$, $1 \leq i \leq n$.

The *likelihood* is the joint pmf of $X = (X_1, \dots, X_n)$, conditional to a given θ, is:

$$p(x \mid \theta) = \theta^{n_1} \, (1 - \theta)^{n - n_1},$$

where $n_1 = \sum_{i=1}^{n} x_i$ is the absolute frequency of ones.

A function of the *sufficient statistic, n_1*.

Marginal pmf of X

$$
\begin{aligned}
p(x) &= \int_{\Theta} p(x \mid \theta)\, p(\theta)\, d\theta \\
&= \int_0^1 \frac{1}{B(\alpha,\beta)}\, t^{\alpha+n_1-1}\,(1-t)^{\beta+n-n_1-1}\, dt \\
&= \frac{1}{B(\alpha,\beta)}\, B(\alpha+n_1, \beta+n-n_1).
\end{aligned}
$$

Prior predictive pdf

$p(x)$ is also called Prior predictive pmf of X.

Prior predictive pdf

$p(x)$ is also called Prior predictive pmf of X.

Why?

Prior predictive pdf

$p(x)$ is also called Prior predictive pmf of X.

Why?

$p(x)$ averages $p(x \mid \theta)$ over all possible θ, each with a relative weight *proportional to the prior* $p(\theta)$.

The Beta-Binomial distribution

For real numbers $\alpha, \beta > 0$, and integer $n > 0$, the pmf:

$$p(k; n, \alpha, \beta) = \binom{n}{k} \cdot \frac{B(\alpha + k, \beta + n - k)}{B(\alpha, \beta)},$$

defines the *Beta-binomial distribution,*

r.v. with support on the set of

nonnegative integers k such that $0 \leq k \leq n$.

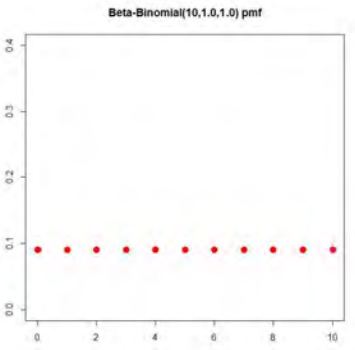
Moments of the Beta-Binomial distribution

For a r.v. $Y \sim \text{Beta-Binom}(n, \alpha, \beta)$

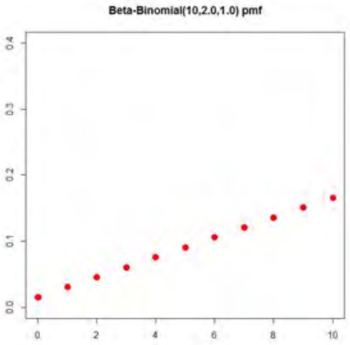
$$\mathrm{E}(Y) = n \cdot \frac{\alpha}{\alpha + \beta},$$

$$\operatorname{var}(Y) = n \cdot \frac{\alpha\,\beta\,(\alpha + \beta + n)}{(\alpha + \beta)^2\,(\alpha + \beta + 1)}.$$

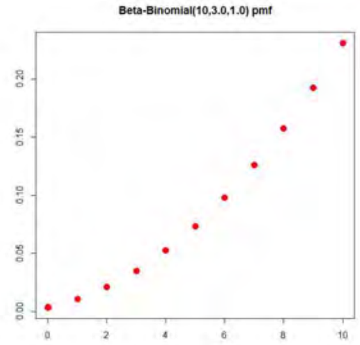
Examples of Beta-Binomial pmf's



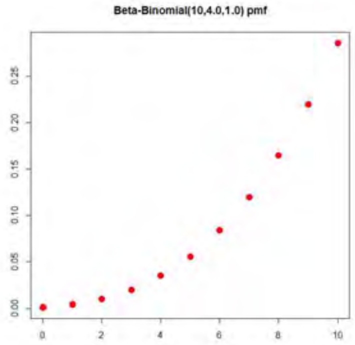
Examples of Beta-Binomial pmf’s



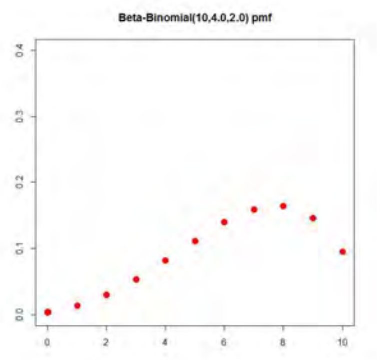
Examples of Beta-Binomial pmf’s



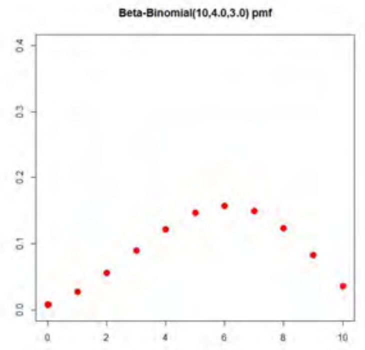
Examples of Beta-Binomial pmf's



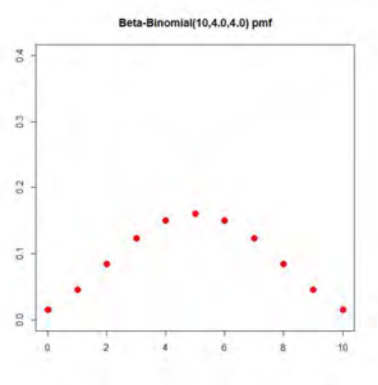
Examples of Beta-Binomial pmf's



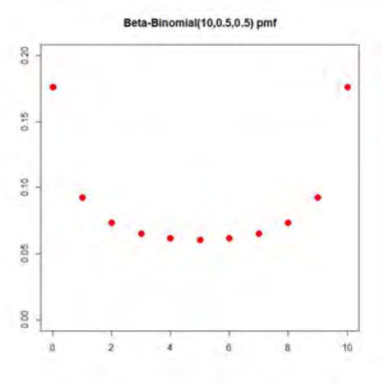
Examples of Beta-Binomial pmf’s



Examples of Beta-Binomial pmf’s



Examples of Beta-Binomial pmf’s



Posterior pdf of θ

Bayes' formula $\rightarrow$ $\quad p(\theta \mid x)$

$$= \frac{p(x \mid \theta)\, p(\theta)}{f(x)}$$

$$= \frac{1}{B(\alpha + n_1, \beta + n - n_1)}\, \theta^{\alpha + n_1 - 1}\, (1 - \theta)^{\beta + n - n_1 - 1}.$$

A conjugate family

The resulting pdf is another Beta distribution,

$$\mathrm{Beta}(\alpha + n_1, \beta + n - n_1).$$

The pair

Bernoulli likelihood / Beta prior

is a conjugate pair.

Posterior expectation of θ

$$E[\theta \mid X = x] \;=\; \frac{\alpha + n_1}{\alpha + \beta + n}.$$

Can be written as a convex combination

$$E[\theta \mid X = x] \;=\; \lambda \cdot \frac{n_1}{n} + (1 - \lambda) \cdot \frac{\alpha}{\alpha + \beta},$$

where $\lambda = \dfrac{n}{\alpha + \beta + n}$.

Posterior expectation of θ

$$\frac{n_1}{n} = \text{empirical probability.}$$

$$\frac{\alpha}{\alpha+\beta} = \text{prior expectation.}$$

Think of prior expectation as the result of a previous experiment, α sucesses out of $\alpha + \beta$ realizations.

Posterior expectation of θ

The coefficient in the convex combination:

$$\lambda = \frac{n}{\alpha + \beta + n}$$

is the ratio of sizes,

actually observed sample

vs. a previous *"virtual"* sample.

Posterior predictive distribution

The Posterior predictive distribution for a new observation $\tilde{x}$, given the observed x, is the average of the pmf $p(x \mid \theta)$ over all possible values of θ, where now relative weights of θ are given by the posterior pdf.

We integrate with respect to θ, the product of the pmf $\operatorname{Binom}(n, \theta)$ times the posterior pdf $\operatorname{Beta}(\alpha + x, \beta + n - x)$.

Posterior predictive distribution

The result is again a Beta-Binomial distribution:

$$p(\tilde{x}) = \frac{1}{B(\alpha + x, \beta + n - x)} \times B(\alpha + x + \tilde{x}, \beta + n - x + \widetilde{n} - \tilde{x}) \binom{\widetilde{n}}{\tilde{x}}.$$

[To allow for the case when the new observation $\tilde{x}$ comes from a different number $\widetilde{n}$ of Bernoulli experiment repetitions, $\tilde{x} \sim \mathrm{Binom}(\widetilde{n}, \theta)$.]

Summary: Beta-Binomial (Bernoulli) model

- Prior distribution of θ: A Beta pdf,
- Prior predictive of x: A Beta-Binomial pdf,
- Posterior of θ, given x: A Beta pdf,
- Posterior predictive of $\tilde{x}$, given x: A Beta-Binomial pdf.

04 - Binomial model 01

Estimating a probability

Which is the least informative prior?

How does choice of prior reflect on the posterior?

With a Bernoulli likelihood, it is not obvious that $\mathrm{Unif}(0,1)$ is “the" Non-Informative Prior (NIP).

Beta priors, plus improper Beta distributions of the form:

$$p(\theta) \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}, \qquad \alpha, \beta \in \mathbb{R}.$$

Zhu, Mu; Lu, Arthur Y. (2004), The Counter-Intuitive Non-informative Prior for the Bernoulli Family, Journal of Statistics Education, ***12*** *(2).*

Useful formulas (1)

With a $\operatorname{Beta}(\alpha, \beta)$ prior pdf, the marginal pmf of x is a Beta-binomial:

$$p(x) = \frac{1}{B(\alpha, \beta)} B(\alpha + n_1, \beta + n - n_1),$$

where $n_1 = \sum_{i=1}^{n} x_i$.

Useful formulas (2)

The expectation and variance of $U \sim \mathrm{Beta}(\alpha, \beta)$ are:

$$\begin{aligned}
\mathrm{E}(U) &= \frac{\alpha}{\alpha+\beta}, \\
\mathrm{var}(U) &= \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}.
\end{aligned}$$

Useful formulas (3)

The posterior pdf of θ, given x:

$$
\begin{aligned}
p(\theta \mid x) &= \frac{p(x \mid \theta) \cdot p(\theta)}{p(x)} \\
&= \frac{1}{B(\alpha + n_1, \beta + n - n_1)}\, \theta^{\alpha + n_1 - 1}\, (1 - \theta)^{\beta + n - n_1 - 1},
\end{aligned}
$$

is a $\mathrm{Beta}(\alpha + n_1, \beta + n - n_1)$ distribution.

Posterior expectation and variance

For the posterior pdf, a $\operatorname{Beta}(\alpha + n_1, \beta + n - n_1)$,

$$
\begin{aligned}
\mathrm{E}(\theta \mid x) &= \frac{\alpha + n_1}{\alpha + \beta + n}, \\
\operatorname{var}(\theta \mid x) &= \frac{(\alpha + n_1)\,(\beta + n - n_1)}{(\alpha + \beta + n)^2\,(\alpha + \beta + n + 1)}.
\end{aligned}
$$

NIP 1: The uniform law

$$p_1(\theta) \sim \mathrm{Unif}[0,1] = \mathrm{Beta}(1,1).$$

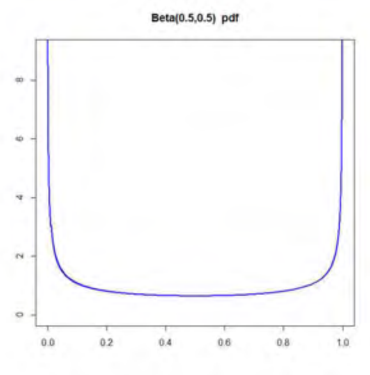
$$\begin{aligned} \mathrm{E}(\theta \mid x) &= \frac{n_1+1}{n+2}, \\ \mathrm{var}(\theta \mid x) &= \frac{(n_1+1)\,(n-n_1+1)}{(n+2)^2\,(n+3)}. \end{aligned}$$

NIP 2: Jeffreys' prior

$$p_2(\theta) \sim \mathrm{Beta}(1/2, 1/2).$$

Drawback is, its appearance is not “non-informative": probability concentrates near 0 and 1.

Probability density function of Jeffreys' prior



NIP 2: Jeffreys' prior

With Jeffreys' prior,

$$\begin{aligned}
\mathrm{E}(\theta \mid x) &= \frac{n_1 + 1/2}{n+1}, \\
\mathrm{var}(\theta \mid x) &= \frac{(n_1 + 1/2)\,(n - n_1 + 1/2)}{(n+1)^2\,(n+2)}.
\end{aligned}$$

The Beta(c, c) subfamily

For the Beta subfamily with $\alpha = \beta = c$, where both Jeffreys' and uniform belong:

$$\begin{aligned} \mathrm{E}(\theta \mid x) &= \frac{n_1 + c}{n + 2c}, \\ \mathrm{var}(\theta \mid x) &= \frac{(n_1 + c)\,(n - n_1 + c)}{(n + 2c)^2\,(n + 2c + 1)}. \end{aligned}$$

The Beta(c, c) subfamily

A Beta(c, c) prior is equivalent to adding $2c$ virtual observations to the sample, c zeros and c ones.

Writing: $N = n + 2c$, $N_1 = n_1 + c$,

$$\mathrm{E}(\theta \mid x) = \frac{N_1}{N}, \qquad \mathrm{var}(\theta \mid x) = \frac{N_1\,(N - N_1)}{N^2\,(N + 1)}.$$

Comparing Jeffreys' and uniform prior

Jeffreys' prior is less influential than the uniform,

It meddles less in the experiment, contributing only one *virtual observation*, evenly distributed between 0 and 1,

The uniform adds two *virtual observations*, one of each.

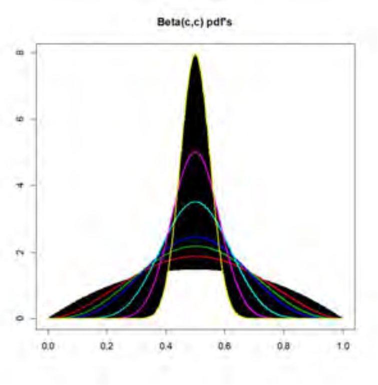
Other Beta(c, c) priors

Within this subfamily,

What happens with a very large or a very small c?

Other Beta(c, c) priors

For $c = 2, 3, 4, 5, 10, 20, 50,$



Other Beta(c, c) priors

If $c \to \infty$, the Beta(c, c) law tends to a degenerate (constant) distribution, with:

$$\mathrm{P}\{\theta = 1/2\} = 1.$$

Then the posterior is this same degenerate law.

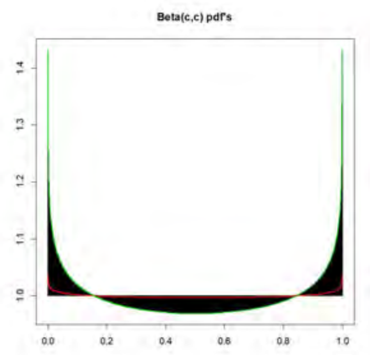
Other Beta(c, c) priors

In agreement with the interpretation above, this is the *dogmatic estimator.*

The *a priori* information is so strong that it overrules any experimental evidence.

Other Beta(c, c) priors

For $c = 1, 0.995, 0.95,$



Other Beta(c, c) priors

In the opposite direction, if $c \to 0$, the less influential prior should be the limit $c = 0$,

$$p(\theta) \propto \theta^{-1} \cdot (1 - \theta)^{-1}, \quad \theta \in (0, 1),$$

for which,

$$\mathrm{E}(\theta \mid x) = \frac{n_1}{n} = f_1, \quad \text{relative frequency of ones,}$$

The classical ML estimator.

Haldane's prior

This Beta(0, 0) pdf can be derived by applying the change of variable formula to the (improper) uniform law:

$$p(\eta) = 1, \quad \eta \in (-\infty, \infty),$$

for the log-odds ratio $\eta = \log\left(\frac{\theta}{1-\theta}\right)$, the natural Bernoulli parameter (as a regular exponential family).

Other Beta(c, c) priors

For $c = 0$,

$$\operatorname{var}(\theta \mid x) = \frac{n_1\,(n - n_1)}{n^2\,(n+1)} = \frac{1}{n+1} f_1 (1 - f_1).$$

Smaller than $\operatorname{var}_\theta(f_1) = \dfrac{1}{n}\,\theta\,(1-\theta)$, the CR bound. !?

Other Beta(c, c) priors

Not a contradiction,

the variance of an estimator $\hat{\theta}(x)$ and

the posterior variance of the parameter θ itself

are entirely different concepts.

Other Beta(c, c) priors

The $c \to 0$ limit, Beta(0, 0), is the discrete law:

$$\mathrm{P}[\theta = 0] = \mathrm{P}[\theta = 1] = 1/2,$$

In a sense, the opposite case to setting $\mathrm{P} = 1$ at $\theta = 0.5$: now there is a maximum indeterminacy between the two extreme possible θ values.

Summary

Jeffreys' prior should appear as reasonably non informative, the *aurea mediocritas* between both “radical"priors.