05 - Binomial model 01

Master in Foundations of Data Science
Bayesian Statistics and Probabilistic Programming
Fall 2018-2019

Josep Fortiana

Wednesday, October 17, 2018

Universitat de Barcelona

05 - Binomial model 01

Reminder: Gamma and Beta

Estimating a probability

Which is the least informative prior?

05 - Binomial model 01

Reminder: Gamma and Beta

Estimating a probability

Which is the least informative prior?

Reminder: Gamma function

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx.$$

For integer a, $\Gamma(a) = (a-1)!$.

In general, $\Gamma(t+1) = t \cdot \Gamma(t)$.

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}.$$

Reminder: Beta function

By definition:

$$B(r,s) = \int_0^1 t^{r-1}\,(1-t)^{s-1}\,dt$$

This function relates to the Gamma function through the identity:

$$B(r,s) = \frac{\Gamma(r)\,\Gamma(s)}{\Gamma(r+s)}.$$

Reminder: Beta function

Useful identities:

$$
\begin{aligned}
B(r,s) &= B(s,r),\\
B(r,s) &= B(r,s+1)+B(r+1,s),\\
B(r+1,s) &= B(r,s)\cdot\frac{r}{r+s},\\
B(r,s+1) &= B(r,s)\cdot\frac{s}{r+s}.
\end{aligned}
$$

Reminder: Beta function

For integer r, s,

$$B(r,s)=\frac{(r-1)!\,(s-1)!}{(r+s-1)!}=\frac{1}{(r+s-1)\times\binom{r+s-2}{r-1}}.$$

Equivalently,

$$\binom{n}{k}=\frac{1}{(n+1)\,B(n-k+1,k+1)}.$$

Reminder: Beta distributions

$U \sim \mathrm{Beta}(\alpha, \beta)$, $\alpha, \beta > 0$, is an absolutely continuous r.v., with support on $[0, 1]$ and pdf:

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)}\, x^{\alpha-1}\, (1-x)^{\beta-1}.$$

Reminder: Beta distributions

U has expectation:

$$\mathrm{E}(U) = \frac{\alpha}{\alpha + \beta},$$

and variance:

$$\mathrm{var}(U) = \frac{\alpha\,\beta}{(\alpha + \beta)^2\,(\alpha + \beta + 1)}.$$

Reminder: Beta distributions

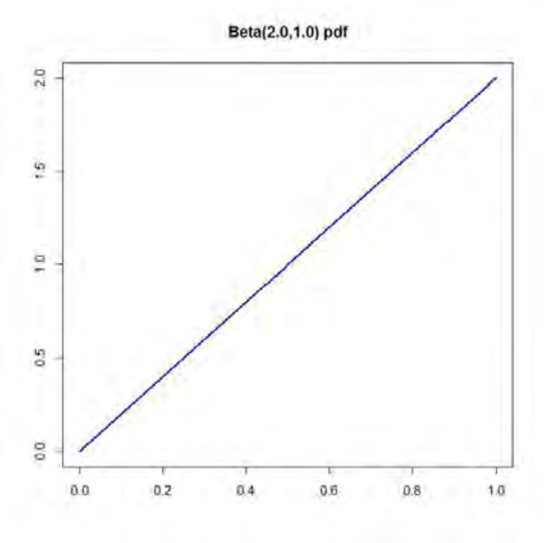
Alternative parameterization, with (θ, p), defined by:

$$
\begin{cases} \theta = \alpha + \beta, \\ p = \dfrac{\alpha}{\alpha + \beta}. \end{cases}
\qquad
\begin{cases} \alpha = \theta \cdot p, \\ \beta = \theta \cdot (1 - p). \end{cases}
$$

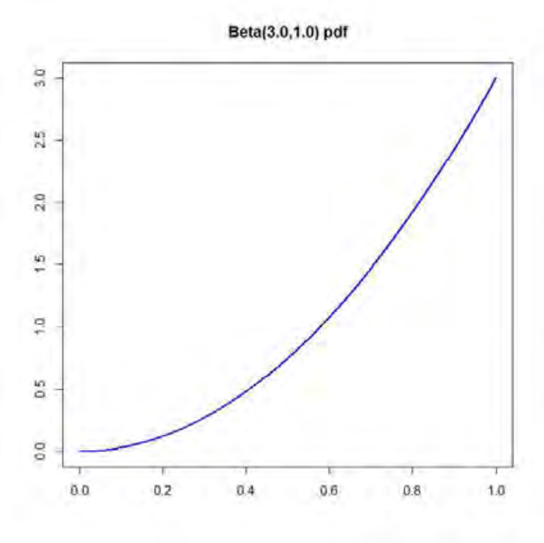
With these parameters:

$$
\mathrm{E}(U) = p, \qquad \operatorname{var}(U) = \frac{p \cdot (1 - p)}{\theta + 1}.
$$

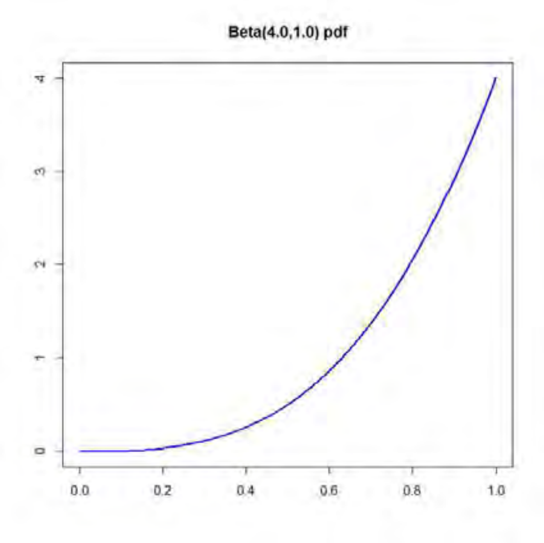
Examples of Beta pdf's with $\alpha, \beta \geq 1$



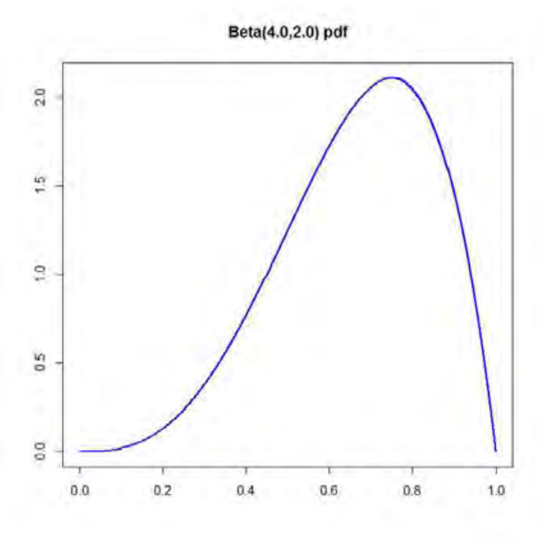
Examples of Beta pdf's with $\alpha, \beta \geq 1$



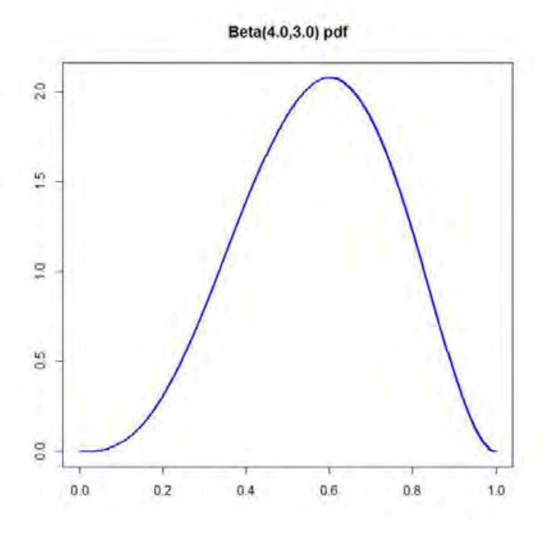
Examples of Beta pdf's with $\alpha, \beta \geq 1$



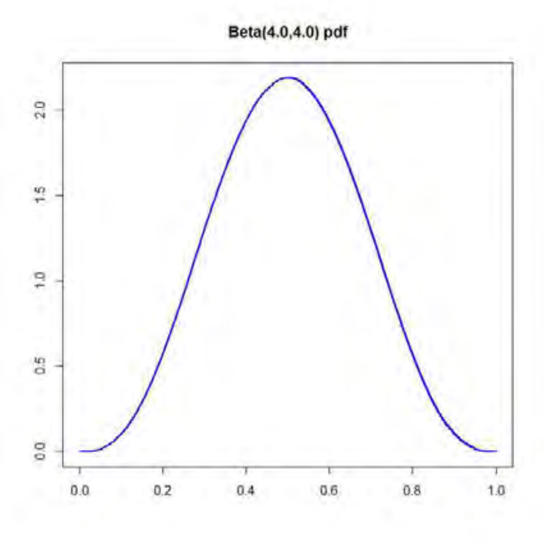
Examples of Beta pdf's with $\alpha, \beta \geq 1$



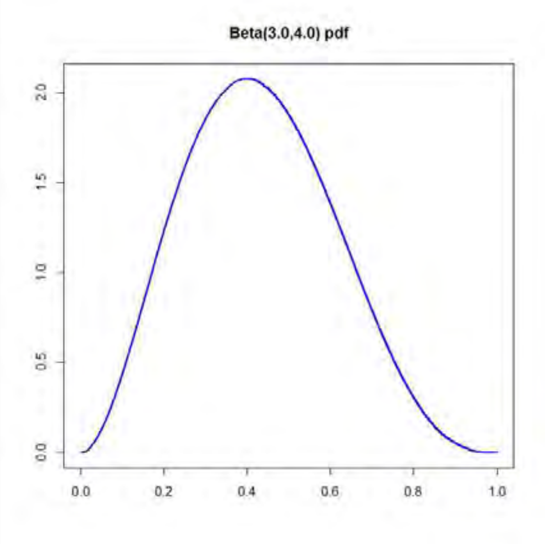
Examples of Beta pdf's with $\alpha, \beta \geq 1$



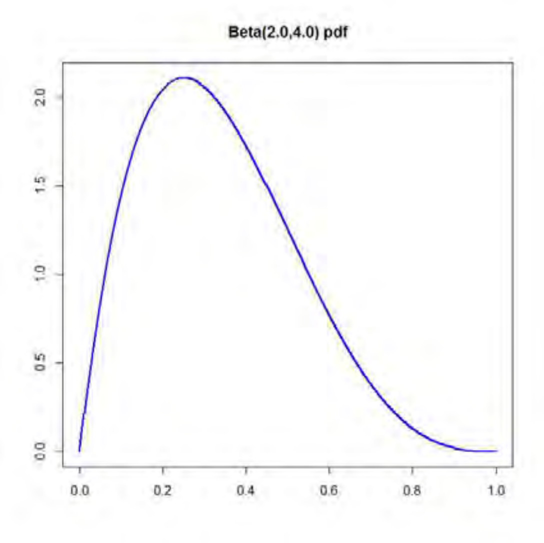
Examples of Beta pdf's with $\alpha, \beta \geq 1$



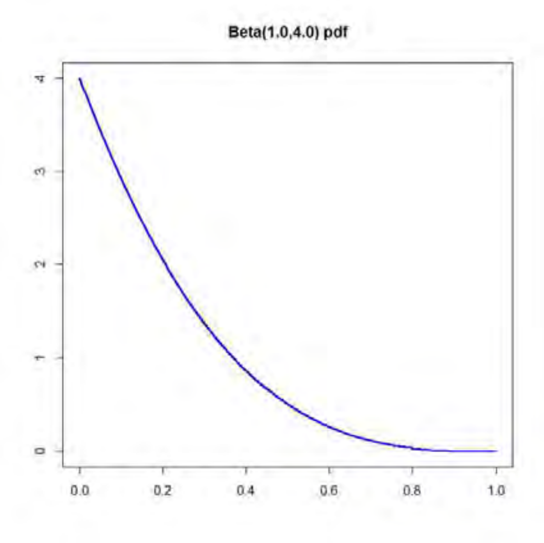
Examples of Beta pdf's with $\alpha, \beta \geq 1$



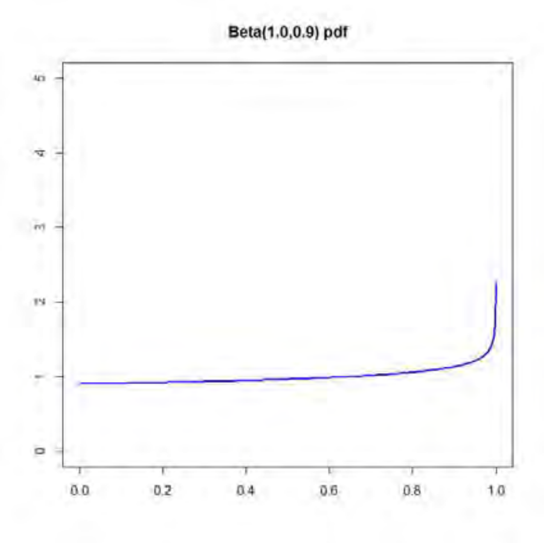
Examples of Beta pdf's with $\alpha, \beta \geq 1$



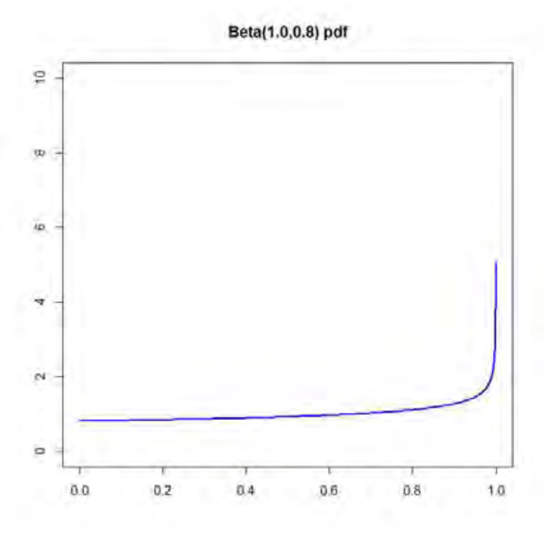
Examples of Beta pdf's with $\alpha, \beta \geq 1$



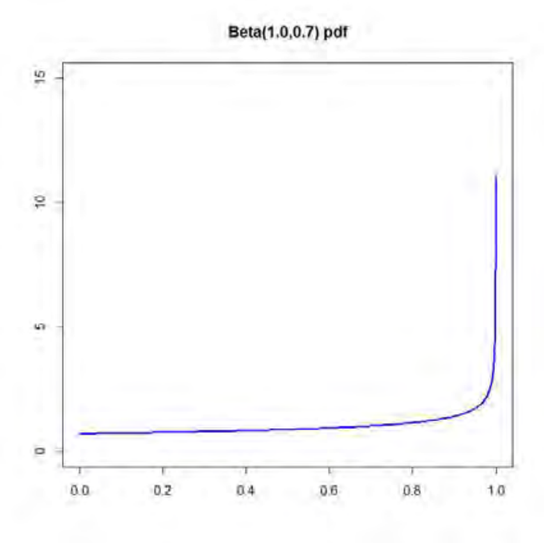
Examples of Beta pdf's with $\alpha, \beta < 1$



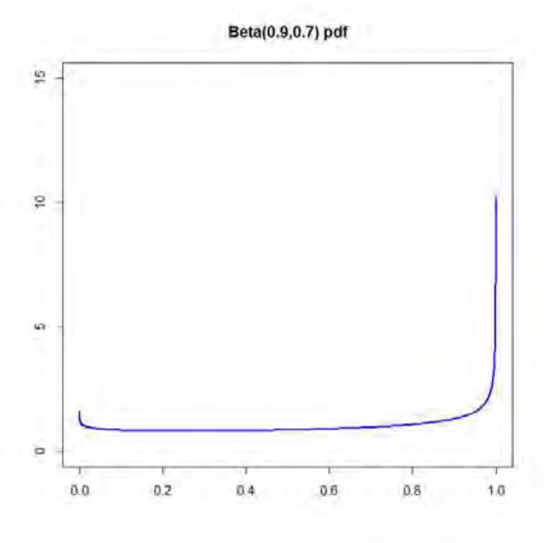
Examples of Beta pdf's with $\alpha, \beta < 1$



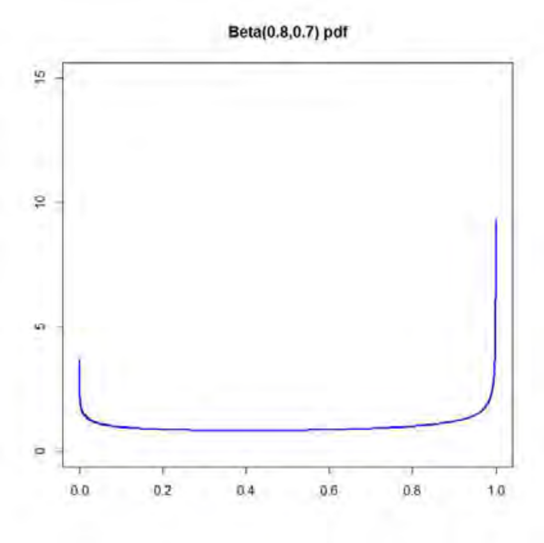
Examples of Beta pdf's with $\alpha, \beta < 1$



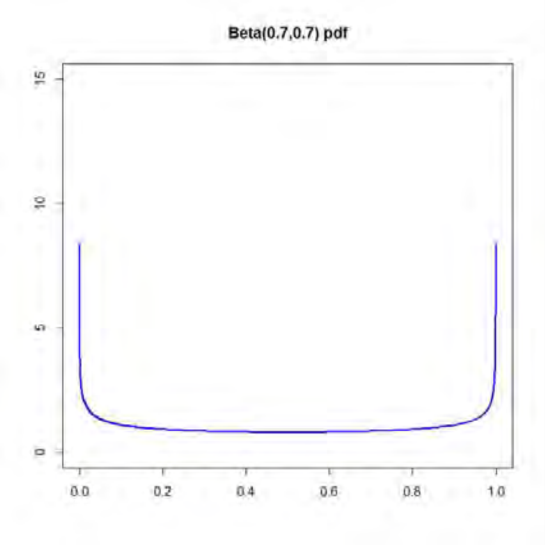
Examples of Beta pdf's with $\alpha, \beta < 1$



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



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



05 - Binomial model 01

Reminder: Gamma and Beta

Estimating a probability

Which is the least informative prior?

Bayesian version of a Bernoulli model

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

We want to estimate the probability $\theta \in \Theta = (0, 1)$.

Prior distribution for θ: if we have no previous information, we can assume a uniform law on $[0, 1]$:

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

A *Non-Informative Prior (NIP)*.

A family of prior distributions

Assume the prior distribution of θ is a beta distribution, $\mathrm{Beta}(\alpha, \beta)$, with pdf:

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

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

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

Likelihood function for the X observations

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

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

$$f(x|\theta) = \theta^{n_1} (1 - \theta)^{n - n_1},$$

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

N.B.: it is a function of the sufficient statistic n_1.

Marginal pmf of X

Integrating, we get the marginal pmf of X:

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

Prior predictive pdf

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

Motivation is:

$f(x)$ averages $f(x|\theta)$ over all possible values of θ, each with its relative weight *according to the prior $h(\theta)$*.

The Beta-Binomial distribution

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

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

defines the *Beta-binomial distribution*,

r.v. with support on the set of

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

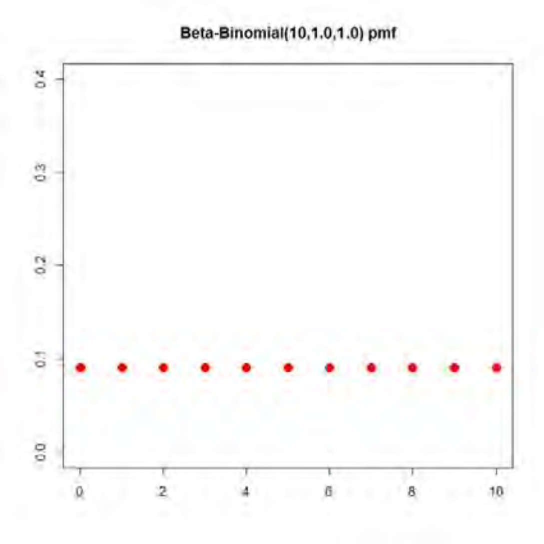
Moments of the Beta-Binomial distribution

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

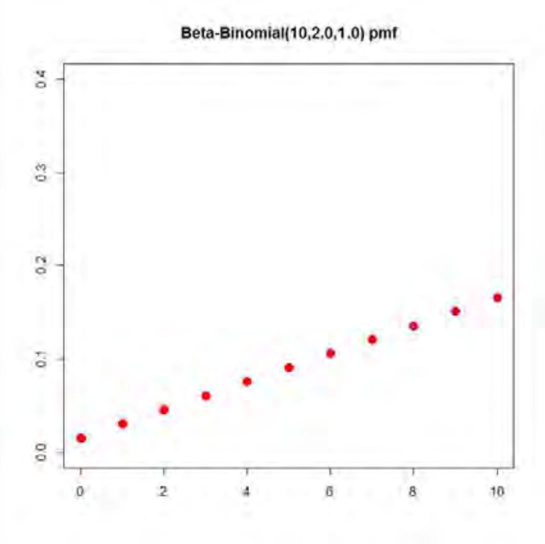
$$\mathrm{E}(Y) = n \cdot \frac{\alpha}{\alpha + \beta},$$

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

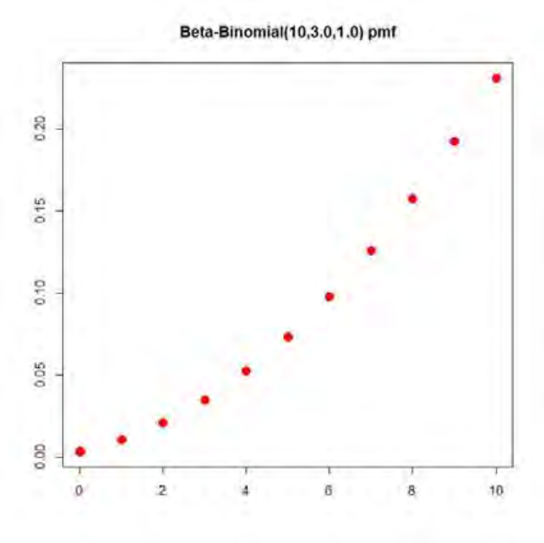
Examples of Beta-Binomial pmf's



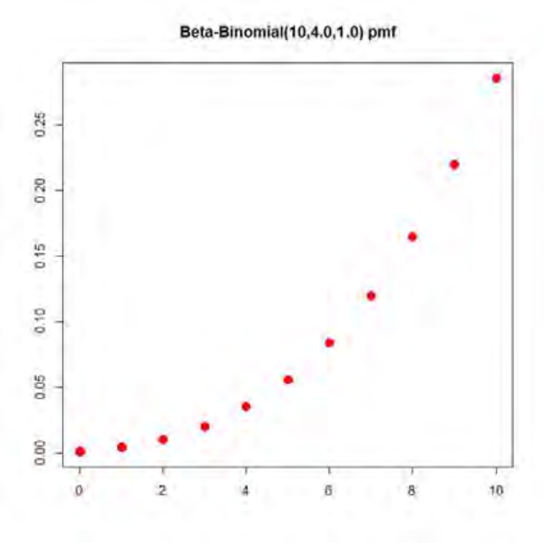
Examples of Beta-Binomial pmf's



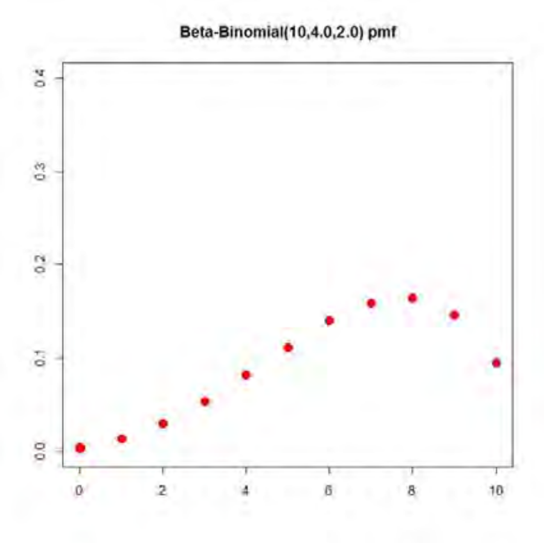
Examples of Beta-Binomial pmf's



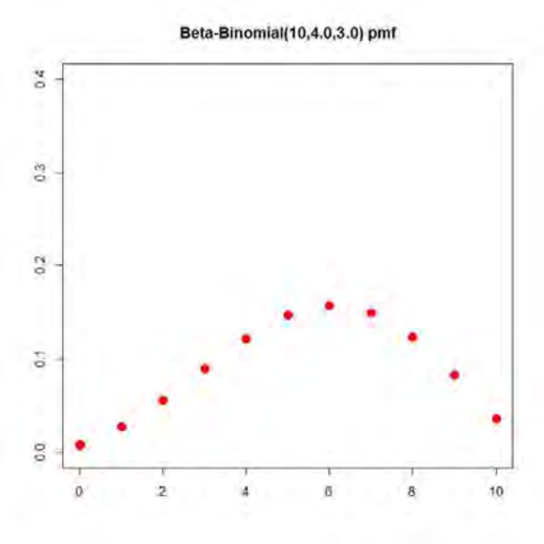
Examples of Beta-Binomial pmf's



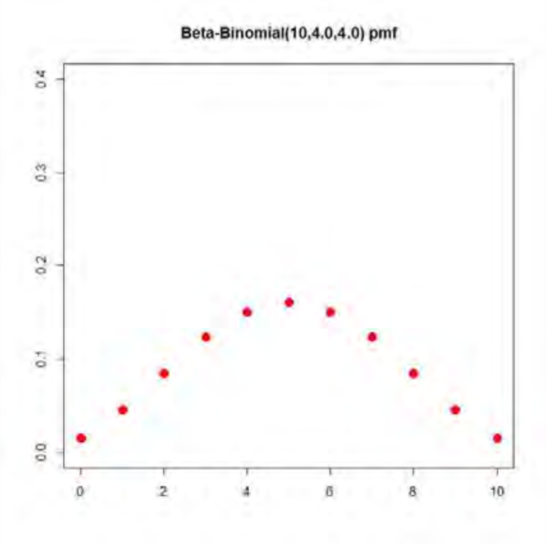
Examples of Beta-Binomial pmf's



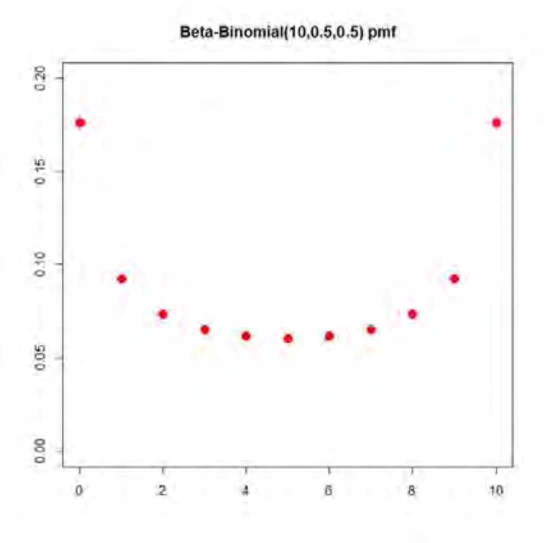
Examples of Beta-Binomial pmf's



Examples of Beta-Binomial pmf's



Examples of Beta-Binomial pmf's



Posterior pdf of θ

Bayes' formula yields: $h(\theta|x) =$

$$
\begin{aligned}
&= \frac{f(x|\theta)\,h(\theta)}{f(x)} \\
&= \frac{1}{B(\alpha+n_1,\beta+n-n_1)}\,\theta^{\alpha+n_1-1}\,(1-\theta)^{\beta+n-n_1-1}.
\end{aligned}
$$

A conjugate family

The resulting pdf is another Beta distribution,

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

The pair formed by a Bernoulli likelihood and a Beta prior is called a conjugate pair.

Posterior expectation of θ

Its expected value is:

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

Interpretation: a convex combination:

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

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

Posterior predictive distribution

The Posterior predictive distribution for a new observation $\tilde{x}$, given the observed x, is the average of the pmf $f(x|\theta)$ 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 $\mathrm{B}(n, \theta)$ times the posterior pdf $\mathsf{Beta}(\alpha + x, \beta + n - x)$.

Posterior predictive distribution

The result is again a Beta-Binomial distribution:

$$
\begin{aligned}
f(\tilde{x}) = & \frac{1}{B(\alpha + x, \beta + n - x)} \\
& \times B(\alpha + x + \tilde{x}, \beta + n - x + \widetilde{n} - \tilde{x}) \binom{\widetilde{n}}{\tilde{x}}.
\end{aligned}
$$

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

Summary: Bernoulli (or Binomial) model with a conjugate prior

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

05 - Binomial model 01

Reminder: Gamma and Beta

Estimating a probability

Which is the least informative prior?

How do changes in the prior reflect on the posterior?

We set a prior pdf in the $\mathsf{Beta}(\alpha, \beta)$ family,

$$h(\theta) \equiv h(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1},$$

We show it is not obvious that $\mathsf{Unif}(0, 1)$ is "the" Non-Informative Prior (NIP) for this problem.

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

Useful formulas (1)

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

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

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

$$B(u, v) = \frac{\Gamma(u)\,\Gamma(v)}{\Gamma(u + v)}, \qquad u, v > 0,$$

is the Beta function.

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:

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

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

Posterior expectation and variance

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

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

Several candidates to NIP for this model

Within the Beta family, extended with improper distributions having the same functional form:

$$h(\theta) \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}.$$

NIP 1: The uniform law

$$h_1(\theta) \sim \mathrm{Unif}[0,1] = \mathrm{Beta}(1,1).$$

With it,

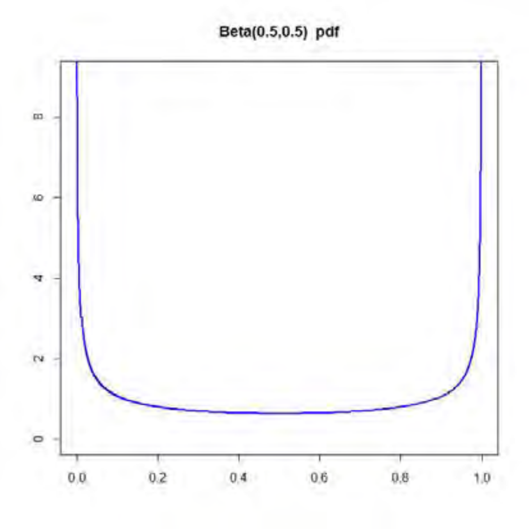
$$\begin{aligned} \mathrm{E}(\theta|x) &= \frac{n_1+1}{n+2}, \\ \mathrm{var}(\theta|x) &= \frac{(n_1+1)\,(n-n_1+1)}{(n+2)^2\,(n+3)}. \end{aligned}$$

NIP 2: Jeffreys' prior

$$h_2(\theta) \sim \mathrm{Beta}(1/2, 1/2).$$

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,

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

The Beta(c, c) subfamily

Consider the subfamily of Beta pdf's with $\alpha = \beta = c$, where both Jeffreys' and uniform belong.

For this subfamily:

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

The Beta(c, c) **subfamily**

Setting a Beta(c, c) prior is equivalent to

adding $2c$ virtual observations to the sample,

c zeros and c ones.

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

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

Comparing Jeffreys' and uniform prior

From this perspective Jeffreys' prior is less influential than the uniform,

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

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

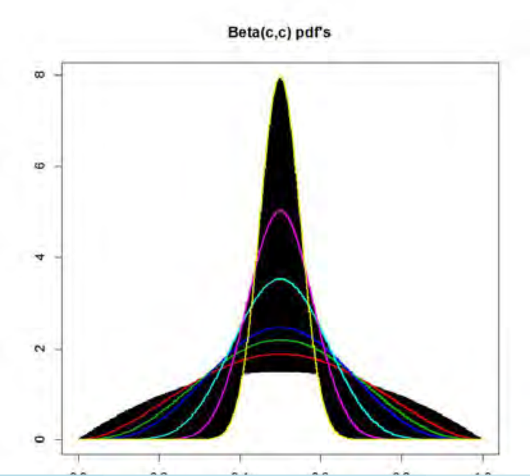
Other Beta(c, c) priors

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 distribution, with:

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

Then the posterior is this same degenerate law.

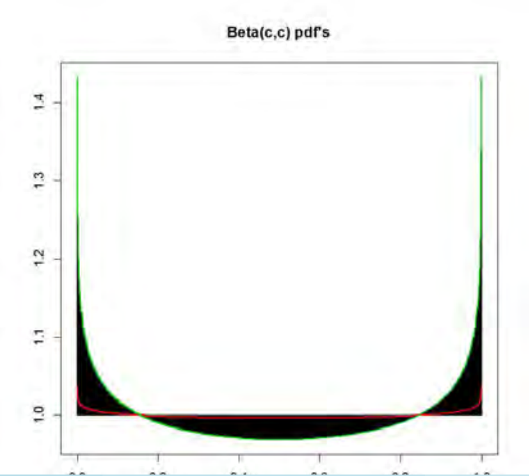
Other Beta(c, c) **priors**

This result is in agreement with the interpretation above, we have the *dogmatic estimator.*

With it, 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 $\mathrm{Beta}(c, c)$ priors

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

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

The classical ML estimator.

Other Beta(c, c) priors

For $c = 0$,

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

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

Not a contradiction, since variance of an estimator $\hat{\theta}(x)$ and posterior variance of the parameter θ itself are entirely different concepts.

Other Beta(c, c) **priors**

The $c \to 0$ limit is the discrete law:

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

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

In the light of these considerations Jeffreys' prior should appear as a reasonably non informative, *aurea mediocritas* between both "radical" priors.

Still another prior: Haldane's prior

Haldane's prior has the improper density:

$$h_4(\theta) = \frac{1}{\theta\,(1-\theta)}, \quad \theta \in (0,1).$$

It is derived by putting an (improper) uniform law on $(-\infty, \infty)$ for the natural parameter:

$$\eta = \log\left(\frac{\theta}{1-\theta}\right),$$

of the model considered as a regular exponential family.

Haldane's prior

h_4 is the result of applying the change of variable formula to the improper density:

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

h_4 can be considered a version of the (inexistent) Beta$(0, 0)$.