Uninformative Priors

Kyle Mann

Sep 22, 2023

Why uninformative prior?

The choice of prior is subjective, which is a common criticism levied against Bayesian methods.

Choosing a prior that is “non-decided” and has little influence over the posterior distribution can protect against such criticisms.

What is your goal:

a) To find the most accurate answer? Then make an informative prior distribution incorporating all information possible.

b) To design an analysis that will convince other people of the validity of your results? Then consider an uninformative prior.

How to choose an “uninformative” prior?

Find a “flat” distribution. For example, for a location parameter that can take on any real value, choose a constant (improper prior) or choose a very flat distribution like $N(0,100000)$.

Problem: “flatness” is dependent on parameterization. If you choose a flat prior, then if you re-parameterize, that prior is no longer flat. How do we make sure that we are using the right scale?

Jeffreys Prior

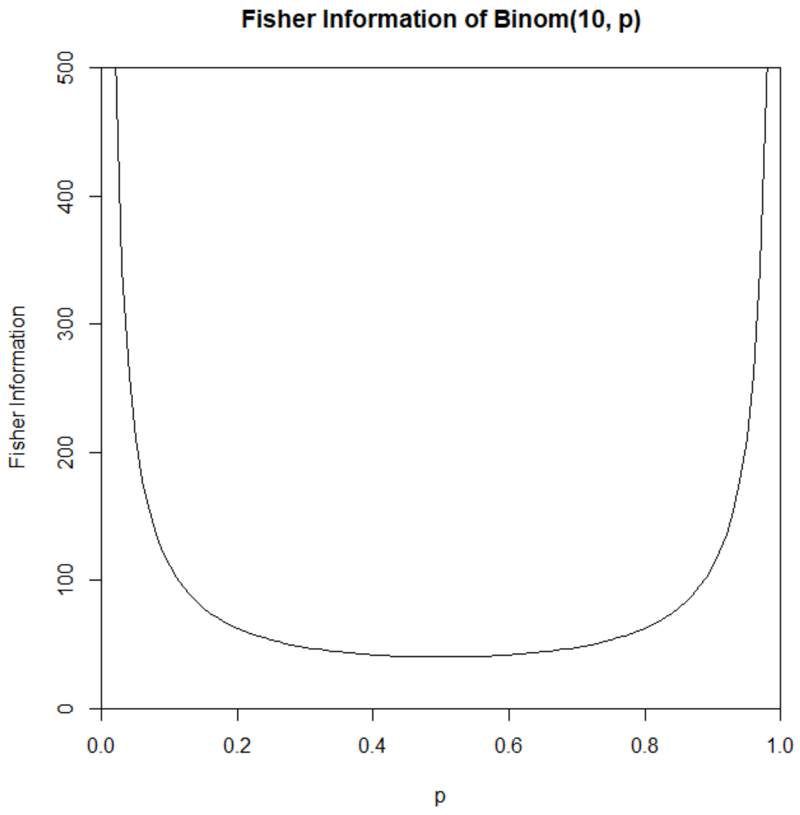
Jeffreys set out to create an uninformative prior distribution whose definition was invariant under re-parameterization.

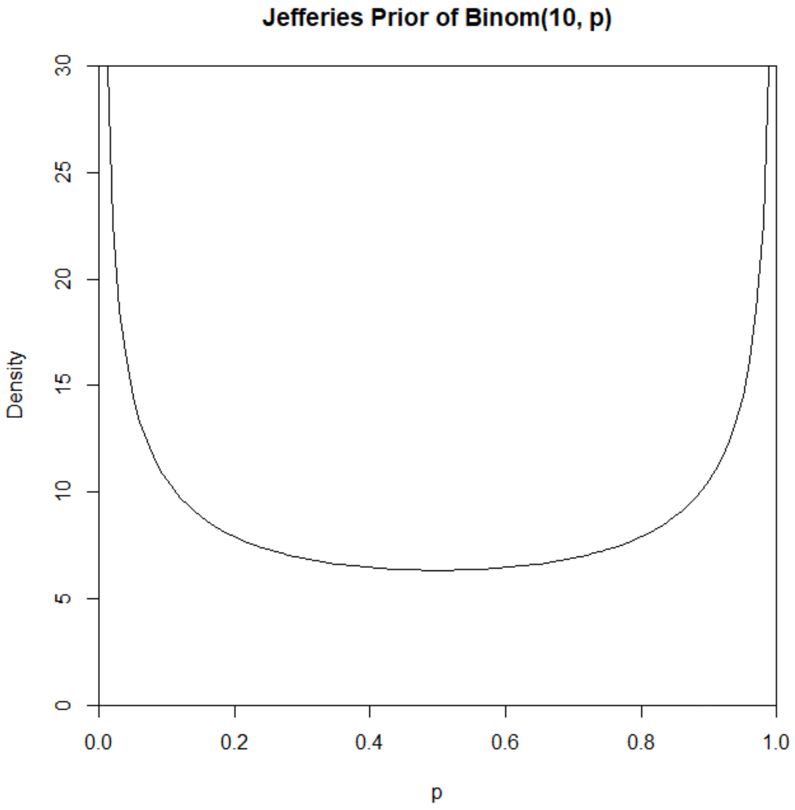
Particularly, if we have a prior distribution $\pi(\theta)$, then if we re-parameterize to $\phi = h(\theta)$, then the prior with respect to ϕ is $\pi(\theta)|h'(\theta)|^{-1}$.

If a prior is defined to be proportional to the square root of the determinant of the Fisher information matrix, then it will be invariant under this transformation.

Jeffreys Prior for Binomial

For example, binomial distribution has Fisher information $\frac{n}{p(1-p)}$.





This gives a $Beta\left(\frac{1}{2},\frac{1}{2}\right)$ distribution.

Jeffreys Prior

Why does the Jeffreys prior give more weights to the parameter values that result in a distribution with higher Fisher information?

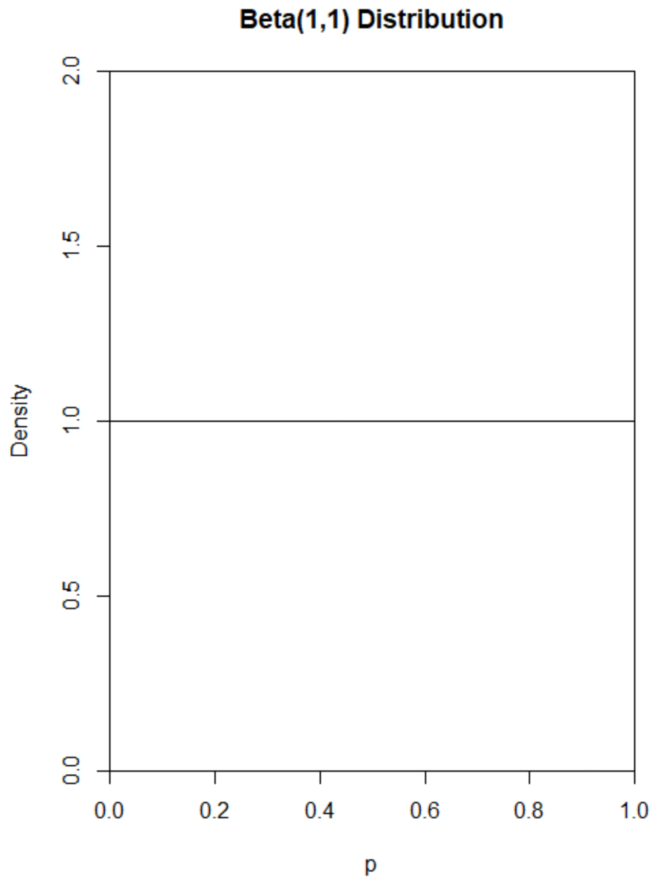
For points of high Fisher information, opposing evidence will be very strong, and so opposing evidence would “quickly” shift the posterior distribution away. For example, in the binomial distribution, If we observe 10 heads in a row, then this will quickly shift the probability away from values such as p=0.01.

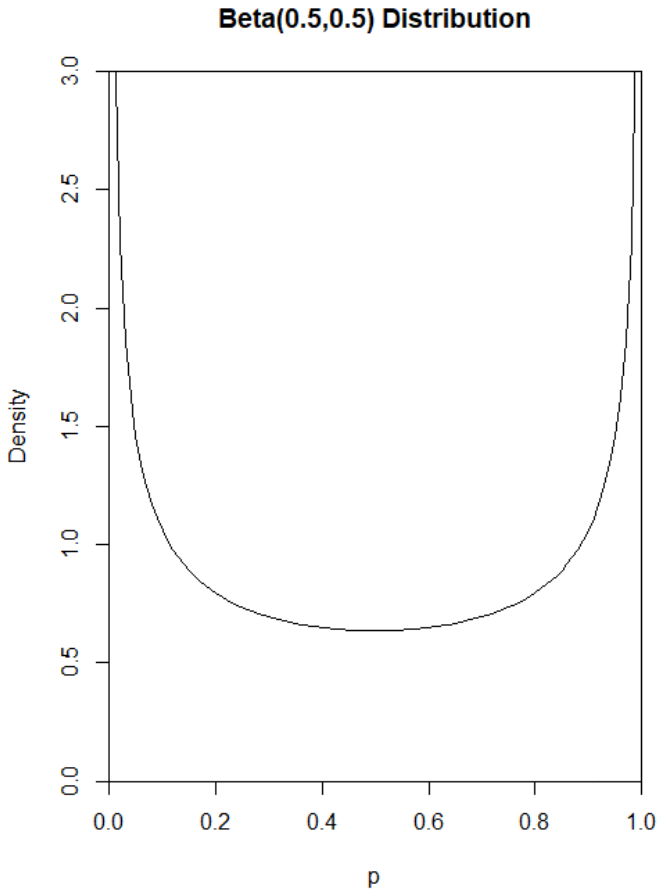
Conjugate families & uninformative priors

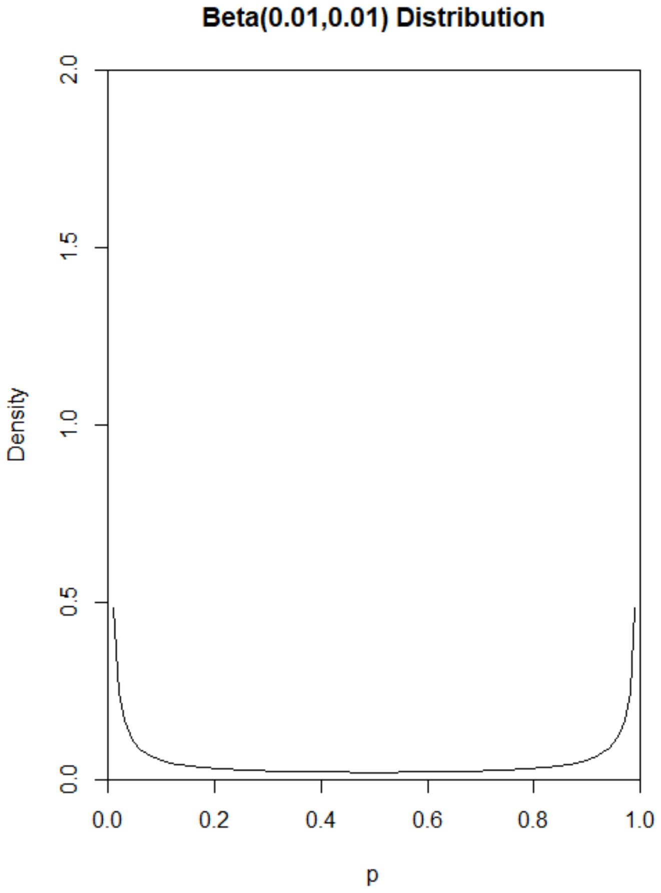
When there is a conjugate family, we can sometimes directly observe the effect of the prior vs the posterior.

If we have a $\mathrm{X}{\sim}Binom(n,p)$ distribution, the conjugate prior is $p{\sim}Beta(\alpha,\beta)$. The posterior distribution is $p{\sim}Beta\big(\alpha+X,\beta+(n-X)\big)$. Remembering $Beta(1,1)$ is uniform and that as the parameters increase, the beta distribution becomes a steeper and steeper hill, we see that the smaller that α and β are, the less they affect the posterior distribution. Thus, $Beta(1,1)$ which is flat in p, $Beta\left(\frac{1}{2},\frac{1}{2}\right)$ the Jeffreys prior, and $Beta(0,0)$ which is flat in the natural exponential family parameter are all uninformative.

Beta-Binomial Uninformative Priors





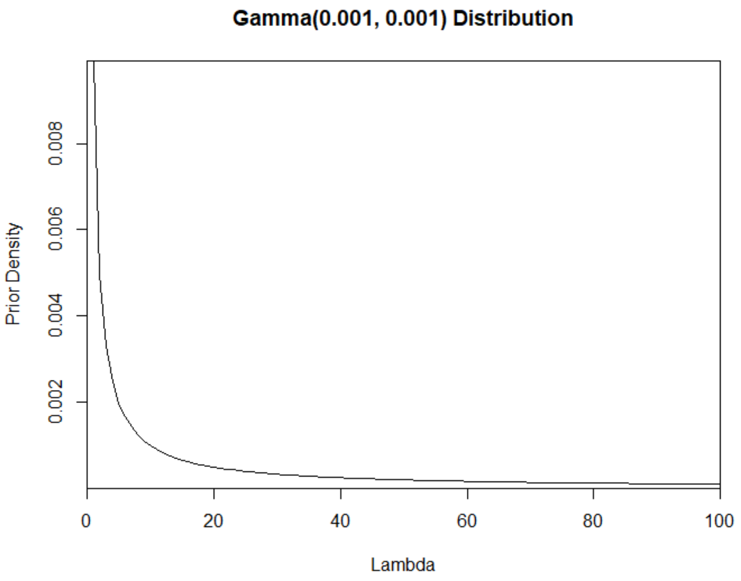


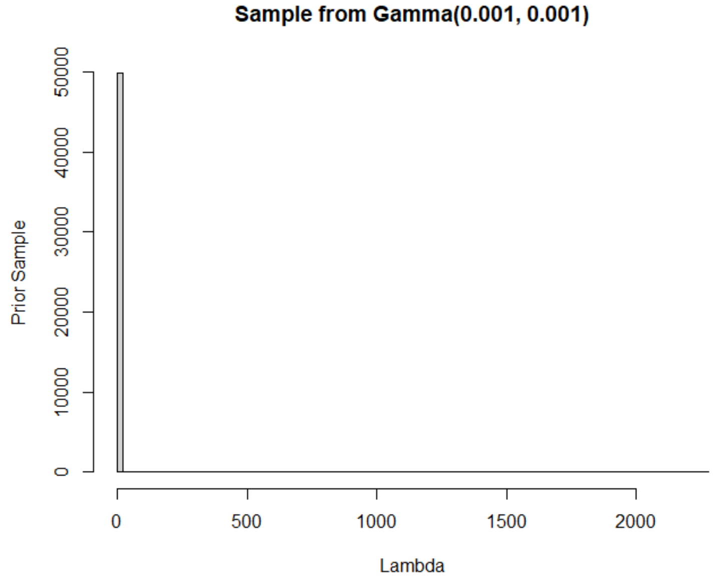
$Beta(0,0)$ is undefined as $\Gamma(0)$ is undefined.

Conjugate families & uninformative priors

Poisson-Gamma

If $X \sim Poisson(\lambda)$, then $\lambda \sim \Gamma(\alpha, \beta)$ is a conjugate prior. The posterior distribution is $\lambda \sim \Gamma(\alpha + n\bar{x}, \beta + n)$. We can see that a small α and β will lead to a posterior dominated by the data.

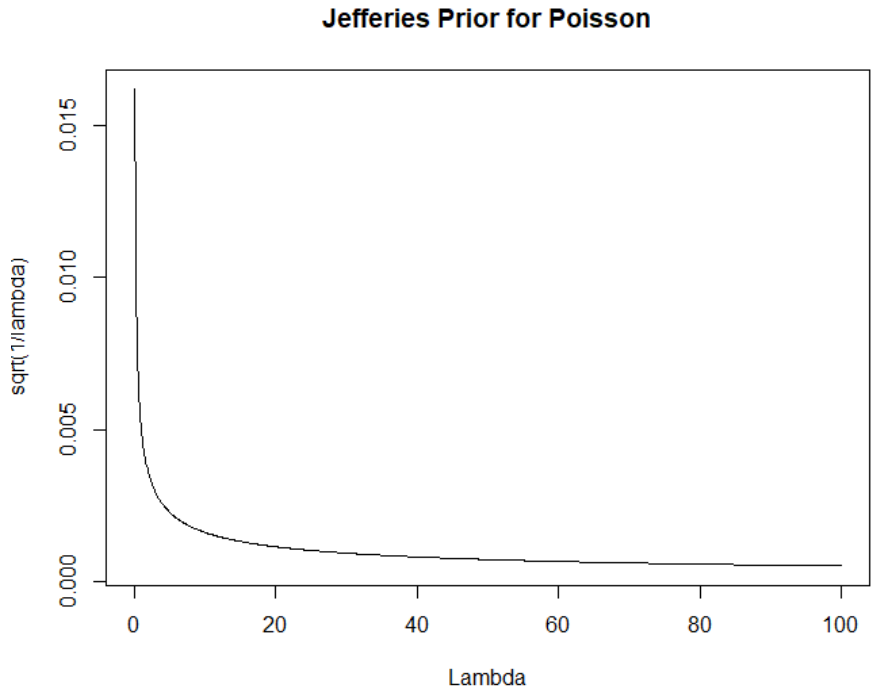




$$\Pr(\lambda < 0.5) = 99\%$$

Jeffreys Prior for Poisson

The Fisher information of a $Poisson(\lambda)$ distribution is $\frac{1}{\lambda}$. Thus, the Jeffreys prior is proportional to $\sqrt{\frac{1}{\lambda}}$. This is plotted:



Conjugate families & uninformative priors

Normal, Unknown Mean

If $X \sim N(\mu, \sigma^2)$ with σ^2 fixed, then the prior $\mu \sim N(\tau, \gamma^2)$ is conjugate. The posterior distribution is

$$\mu \sim N\left(\frac{\sigma^2}{\sigma^2 + n\gamma^2}\tau + \frac{n\gamma^2}{\sigma^2 + n\gamma^2}\bar{X}, \frac{\sigma^2\gamma^2}{\sigma^2 + n\gamma^2}\right)$$

This is a weighted sum of τ and $\bar{X}$. We can see that the larger the prior variance γ^2 is, the more the sum is weighted toward the data rather than the prior.

Jeffreys Prior for Normal Unknown Mean

The Fisher information is $\frac{1}{\sigma^2}$. This is a constant in μ, so an improper constant prior distribution is the Jeffreys prior.

Conjugate families & uninformative priors

Normal, Unknown Variance

If $X \sim N(\mu, \sigma^2)$ with μ fixed, then the prior $\sigma^2 \sim (Scaled\chi^2)^{-1}(\eta, s_0^2)$ is conjugate. The posterior distribution is

$$\sigma^2 \sim (Scaled\chi^2)^{-1}(\mathrm{\eta} + n, \eta s_0^2 + nS^2)$$

We can see that a small η and small s_0^2 will lead to a posterior dominated by the data.

Jeffreys Prior for Normal Unknown Variance

The Fisher information is $\frac{1}{2\sigma^4}$. The Jeffreys prior is thus the improper prior that is proportional to $\frac{1}{\sigma^2}$ in σ^2.

Note that $\pi(\sigma^2) = \frac{1}{\sigma^2}$ is equal to $\pi(\sigma) = \frac{1}{\sigma}$!

Conjugate families & uninformative priors

Normal, Unknown Mean & Variance

If $X \sim N(\mu, \sigma^2)$, then the "$Normal(Scaled\chi^2)^{-1}(\tau, \eta, \gamma^2, s_0^2)$ prior is conjugate:

$$\sigma^2 \sim (Scaled\chi^2)^{-1}(\eta, \gamma^2), \qquad \mu | \sigma^2 \sim N\left(\tau, \frac{\sigma^2}{\eta}\right)$$

The posterior distribution is

$$Normal(Scaled\chi^2)^{-1}\left(\frac{\eta\tau + n\bar{X}}{\eta + n}, \eta + n, \frac{1}{s_0^2 + n}\left[(n-1)S^2 + s_0^2\gamma^2 + \frac{n\eta}{n+\eta}(\bar{X} - \tau)^2\right], s_0^2 + n\right)$$

We can see that a small η, small s_0^2, and small γ^2 will lead to a posterior dominated by the data.

Jeffreys Prior for Normal

The Fisher information matrix is $\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$. The determinant is $\frac{1}{2\sigma^6}$, and thus the Jeffreys prior is constant in μ and proportional to $\sqrt{\frac{1}{\sigma^6}} = \frac{1}{\sigma^3}$ in σ^2.

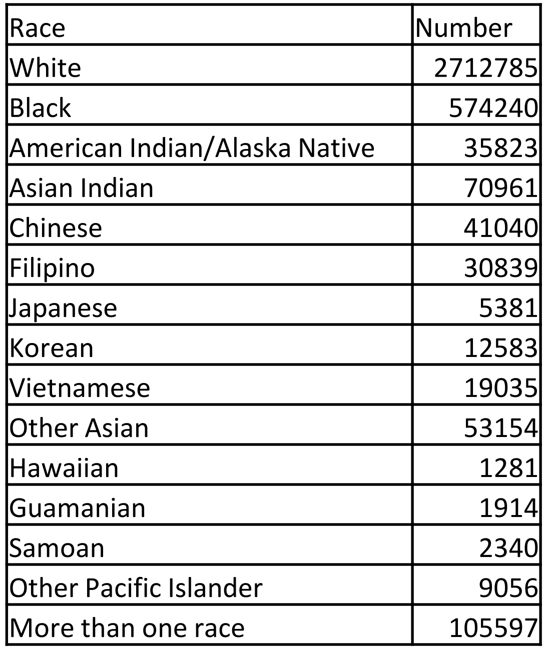
Note that $\pi(\sigma^2) = \frac{1}{\sigma^3}$ is equal to $\pi(\sigma) = \frac{1}{\sigma^2}$!

Larger & More Complex Models

For larger and more complex models, there usually do not exist conjugate families, and the Jeffreys prior may be harder to find. Further, Gelman notes in BDA that the Jeffreys prior can produce counter-intuitive results in high parameter-dimension cases.

- A common practice in hierarchical models is to assign each hyperparameter a prior that is individually uninformative (without considering the joint distribution).

Case Study – 2022 Newborns by race

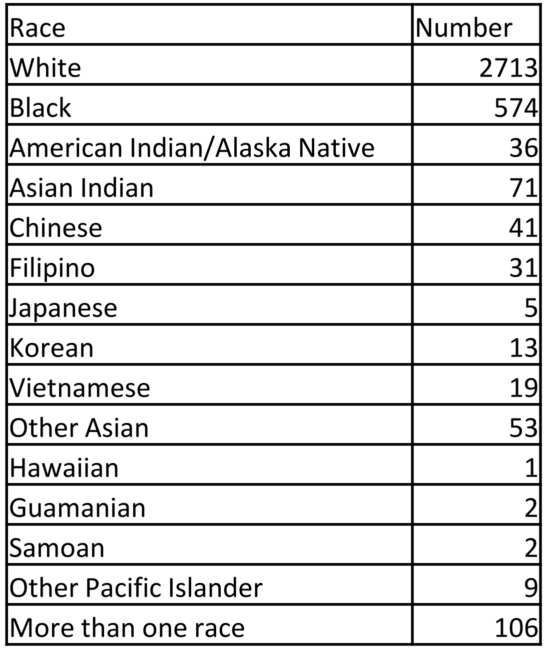


Race	Number
White	2712785
Black	574240
American Indian/Alaska Native	35823
Asian Indian	70961
Chinese	41040
Filipino	30839
Japanese	5381
Korean	12583
Vietnamese	19035
Other Asian	53154
Hawaiian	1281
Guamanian	1914
Samoan	2340
Other Pacific Islander	9056
More than one race	105597

This is complete data!
We do not need to do
any statistical modeling!

Case Study – 2022 Newborns by race

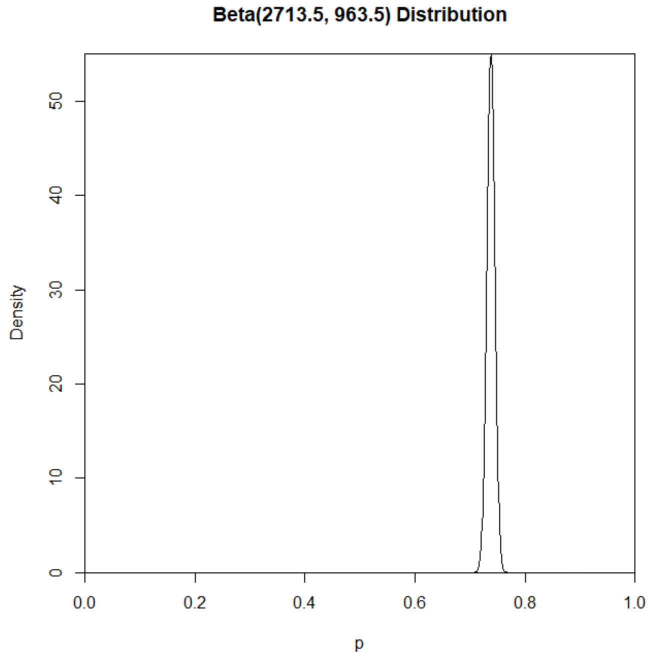
For illustration, let’s pretend that we could only take a limited sample, say 1 in 1000 people. Dividing the numbers by 1000 and rounding to the nearest integer:



Race	Number
White	2713
Black	574
American Indian/Alaska Native	36
Asian Indian	71
Chinese	41
Filipino	31
Japanese	5
Korean	13
Vietnamese	19
Other Asian	53
Hawaiian	1
Guamanian	2
Samoan	2
Other Pacific Islander	9
More than one race	106

Modeling the % White

We can use a Beta-Binomial model with the Jeffreys prior of $p \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$ for the percent white p. We observe 2713/3676 (73.8%) white newborns and 963/3676 (26.2%) non-white newborns. The posterior distribution is $Beta(2713.5, 963.5)$.



Modeling the % of each race

The conjugate prior for a multinomial distribution is a Dirichlet distribution.

Note: In Dirichlet distribution, X_i is independent of $X^{(-i)} := \left\{\frac{X_j}{1-X_i} | j \neq i\right\}$. Furthermore, the marginal distribution of X_i is:

$$X_i \sim Beta\left(\alpha_i, \sum_{j\neq i} \alpha_j\right)$$

The Jeffreys prior is $p \sim Dirichlet\left(\frac{1}{2}, \dots, \frac{1}{2}\right)$. The posterior distribution is:

$$p \sim Dirichlet(2713.5,574.5,36.5 \dots, 106.5)$$

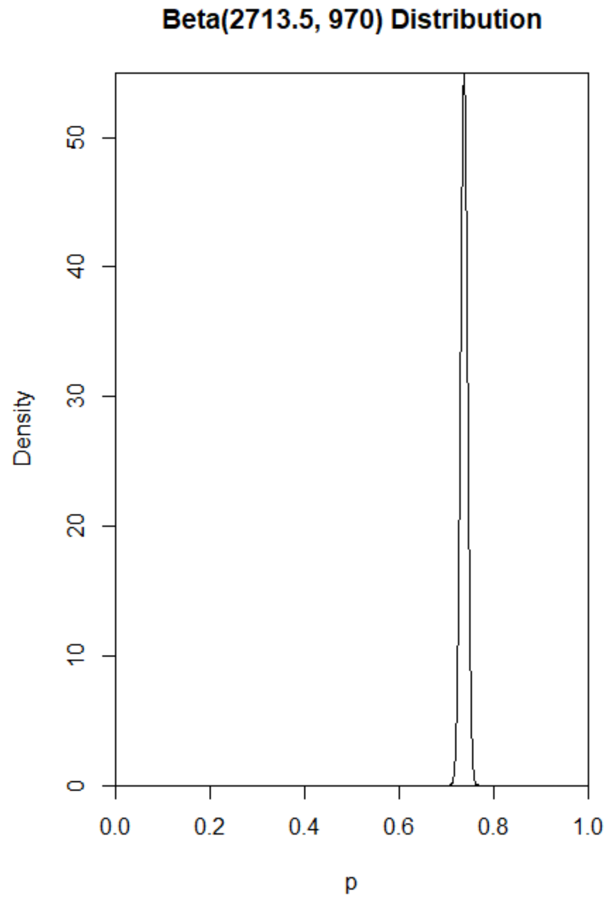
Modeling the % of each race

The posterior distribution is:
$p \sim Dirichlet(2713.5,574.5,36.5 \ldots, 106.5)$

The marginal distribution for the probability of a newborn being white is now:

$$p_{white} \sim Beta\left(\frac{1}{2} + 2713, \frac{14}{2} + 963\right)$$

This differs from previous model. In that it has $\frac{14}{2}$ instead of $\frac{1}{2}$. But we have so much data, posterior is hardly changed:



Note: A $Dirichlet(0, \ldots, 0)$ model would result in the same posterior.

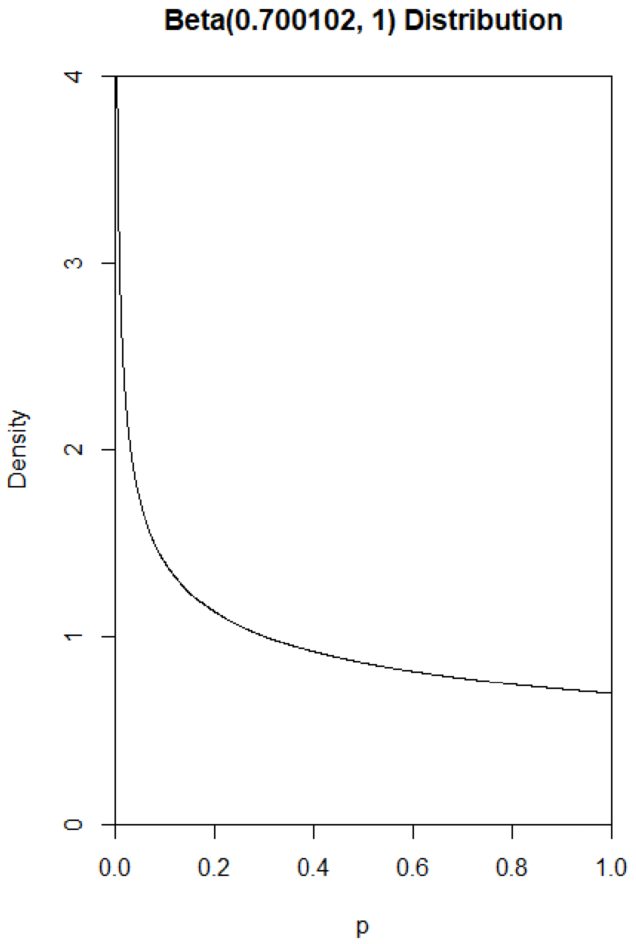
Other considerations

- If we have a success criterion based on the posterior probability, say we will declare success if the posterior probability of a parameter being >0 is 90% or higher, then assess the probability under the prior distribution. If it is too high, then prior can be seen to be influencing the results.
- The frequentist operating characteristics can be assessed by simulating data under various scenarios. The effect of prior choice on the frequentist Type 1 error can be assessed.

Real World Examples – Pfizer Vaccine Trial

Parameter is $\theta = \frac{1-VE}{2-VE}$, where Vaccine Efficacy is $VE = 100(1 - IRR)$ and infection rate ratio is $IRR = \frac{IR_{treat}}{IR_{control}}$. $\theta \in [0,1]$, values less than 0.5 indicate efficacy and values greater than 0.5 indicate harm.

The prior was $Beta(0.700102,1)$. It is centered at $\theta = 0.4118$ and has 96% CI $(0.005,0.964)$.



Question: Why not use 2 separate Beta Binomial models?

Real World Examples – Pfizer Vaccine Trial

At interim analyses, stop for success if $P(VE > 30\%|data) > 0.995$ and at the final analysis declare success if $P(VE > 30\%|data) > 0.986$. At the 93 cases IA, there were 4 in treatment and 90 in control, showing efficacy.

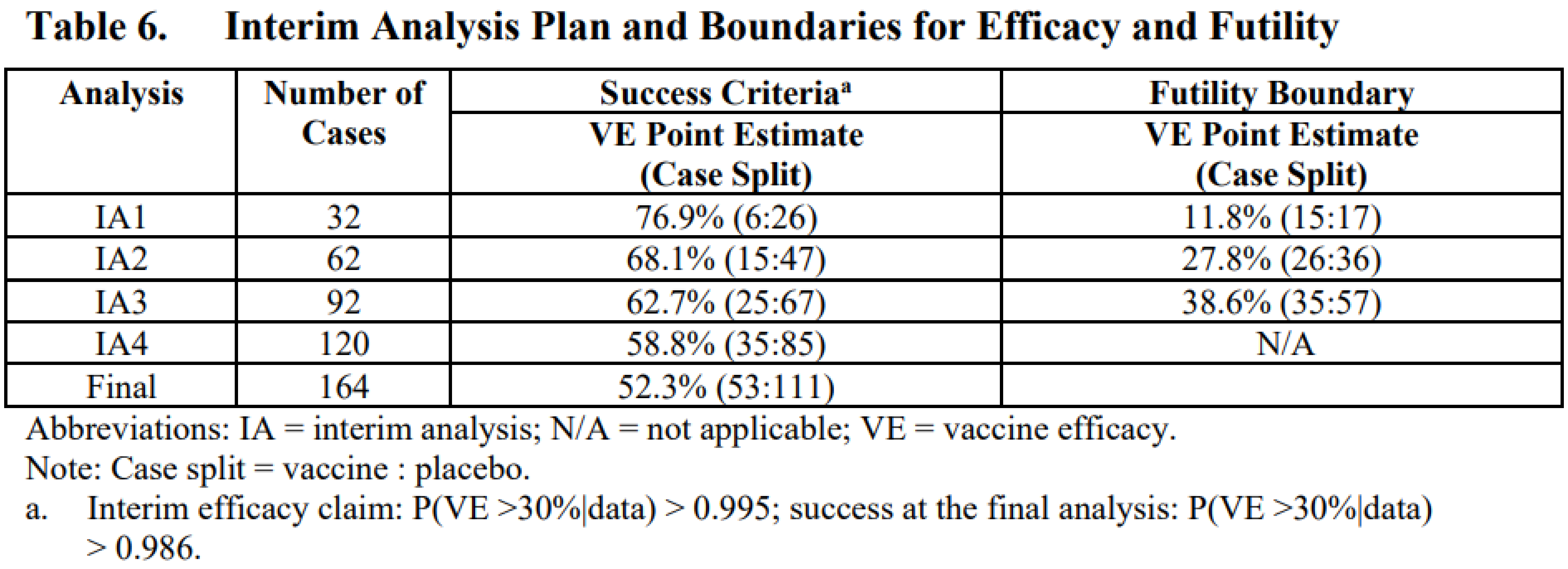


Table 6. Interim Analysis Plan and Boundaries for Efficacy and Futility

Analysis	Number of Cases	Success Criteria[a]	Futility Boundary
		VE Point Estimate (Case Split)	VE Point Estimate (Case Split)
IA1	32	76.9% (6:26)	11.8% (15:17)
IA2	62	68.1% (15:47)	27.8% (26:36)
IA3	92	62.7% (25:67)	38.6% (35:57)
IA4	120	58.8% (35:85)	N/A
Final	164	52.3% (53:111)	

Abbreviations: IA = interim analysis; N/A = not applicable; VE = vaccine efficacy.

Note: Case split = vaccine : placebo.

[a] Interim efficacy claim: P(VE >30%|data) > 0.995; success at the final analysis: P(VE >30%|data) > 0.986.

Real World Examples – GSK2798745

Trial testing whether GSK2798745 a TRPV4 inhibitor could treat chronic cough. The endpoint was the log-transformed daytime cough counts following 7 days of dosing. Let r be the ratio of this log transformed count in the treatment group vs the control group. Planned that n=24 with an interim analysis at n=12. Crossover design, so each patient gets both treatment & control.

Success criterion: $PR(r < 0.7) > 70\%$

Futility criterion: $PR(r < 0.7) < 30\%$

Real World Examples – GSK2798745

Building the model: Let $y_i = \log(count_i)$

Likelihood: Assumed that $\boldsymbol{y} \sim \boldsymbol{\beta X} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is block diagonal with a 2x2 block for each subject across the 2 crossover periods. This is a Bayesian version of a mixed effects model.

Prior: Fixed effects: $\beta_i \sim N(0{,}100000)$

Covariance matrix: Assumed to be unstructured with each block $\begin{pmatrix} \sigma_1^2 & \sigma_{21} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \sim InvWishart$ (conjugate prior for multivariate normal)

If unstructured gives poor fit, the switch to AR(1), with $\begin{pmatrix} \sigma^2 & \sigma^2\phi \\ \sigma^2\phi & \sigma^2 \end{pmatrix}$, $\phi \sim U(-1{,}1)$, $\sigma^2 \sim invGamma(0.0001, 0.0001)$