Introduction to Bayesian Methods- 2

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Likelihood

Prior distribution

Posterior distribution

$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

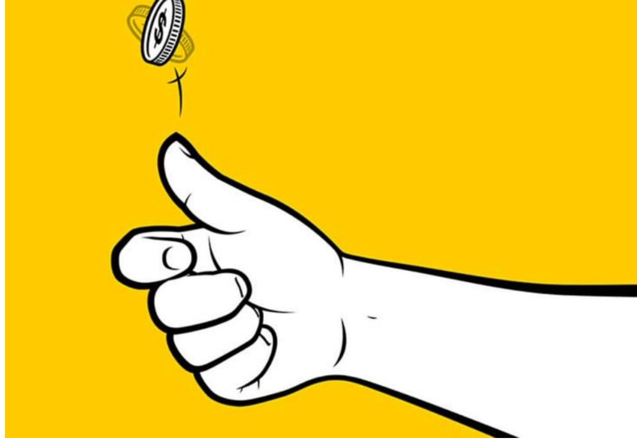
Evidence

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_j P(D|H_j)P(H_j)} P(H_k)$$

$$p(\theta|D,I) = \frac{P(D|\theta,I)}{\int_\Theta P(D|\theta',I)p(\theta'|I)d\theta} p(\theta|I)$$

MAP estimates

Consider the following sequence of coin tosses



`H, T, T, T, H, H, T, T, H, H, T, H, H, H, H, T, H, H, H, T`

(12 heads, 8 tails)

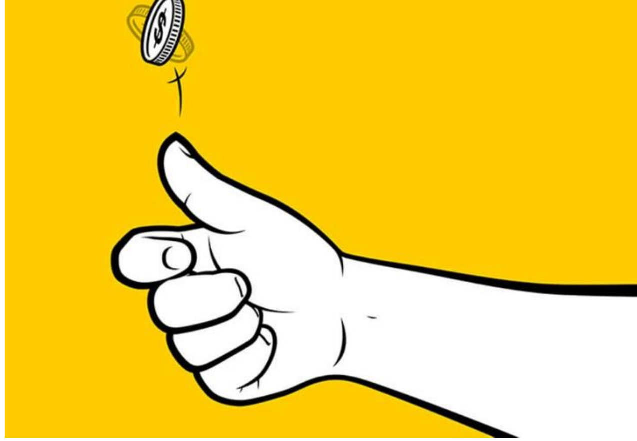
Is this an unbiased coin?

What about the following one?

`H, T, H, H, H, T, H, H, T, H, H, H, H, H, T, T, H, T, T, H`

(13 heads, 7 tails)

Consider the following sequence of coin tosses



`H, T, T, T, H, H, T, T, H, H, T, H, H, H, H, T, H, H, H, T`

(12 heads, 8 tails)

Is this an unbiased coin?

Frequentist answer: sample average is 0.6 instead of 0.5, which is just a little less that one standard deviation (≈ 0.11) away from the mean for an unbiased coin

What about the following one?

`H, T, H, H, H, T, H, H, T, H, H, H, H, H, T, T, H, T, T, H`

(13 heads, 7 tails)

Frequentist answer: sample average is 0.65 instead of 0.5, which is just about 1.36 standard deviations away from the mean for an unbiased coin and is more likely to point to a biased coin

Example of Bayesian inference:

estimate of the (probability) parameter of the binomial distribution

$$P(n|\theta, N) = \binom{N}{n}(1-\theta)^{N-n}\theta^n$$

this is the parameter that we want to infer from data

$$p(\theta|n, N) = \frac{P(n|\theta, N)}{\int_0^1 P(n|\theta', N)p(\theta')d\theta'}\, p(\theta)$$

uniform distribution: the least informative prior

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)'^{N-n}\theta'^n\; p(\theta')d\theta'}\, p(\theta) = \frac{(1-\theta)^{N-n}\theta^n}{\int_0^1 (1-\theta)'^{N-n}\theta'^n\; d\theta'}$$

the final result is a beta distribution

$$p(\theta|n,N) = \frac{(1-\theta)^{N-n}\theta^{n}}{\int_0^1 (1-\theta)'^{N-n}\theta'^{n}\; d\theta'} = \frac{(1-\theta)^{N-n}\theta^{n}}{B(n+1, N-n+1)}$$

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

beta function

$$\mathrm{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

beta pdf

➡ $$p(\theta|n,N) = \mathrm{Beta}(n+1, N-n+1)$$

Mathematical digression: the connection between gamma and beta function

$$\Gamma(a)\Gamma(b) = \int_0^\infty s^{a-1}e^{-s}ds \int_0^\infty t^{b-1}e^{-t}dt$$

$$s = x^2,\ t = y^2 \quad \Rightarrow \quad \Gamma(a)\Gamma(b) = 4\int_0^\infty x^{2a-1}e^{-x^2}dx \int_0^\infty y^{2b-1}e^{-y^2}dy$$

$$x = r\cos\theta,\ y = r\sin\theta \quad \Rightarrow \quad \Gamma(a)\Gamma(b) = 4\int_0^\infty r^{2a+2b-1}e^{-r^2}dr \int_0^{\pi/2} \cos^{2a-1}\theta \sin^{2b-1}\theta d\theta$$

$$(t = \cos^2\theta,\ dt = -2\cos\theta\sin\theta d\theta) \qquad = \Gamma(a+b)\left(2\int_0^{\pi/2}\cos^{2a-1}\theta\sin^{2b-1}\theta d\theta\right)$$

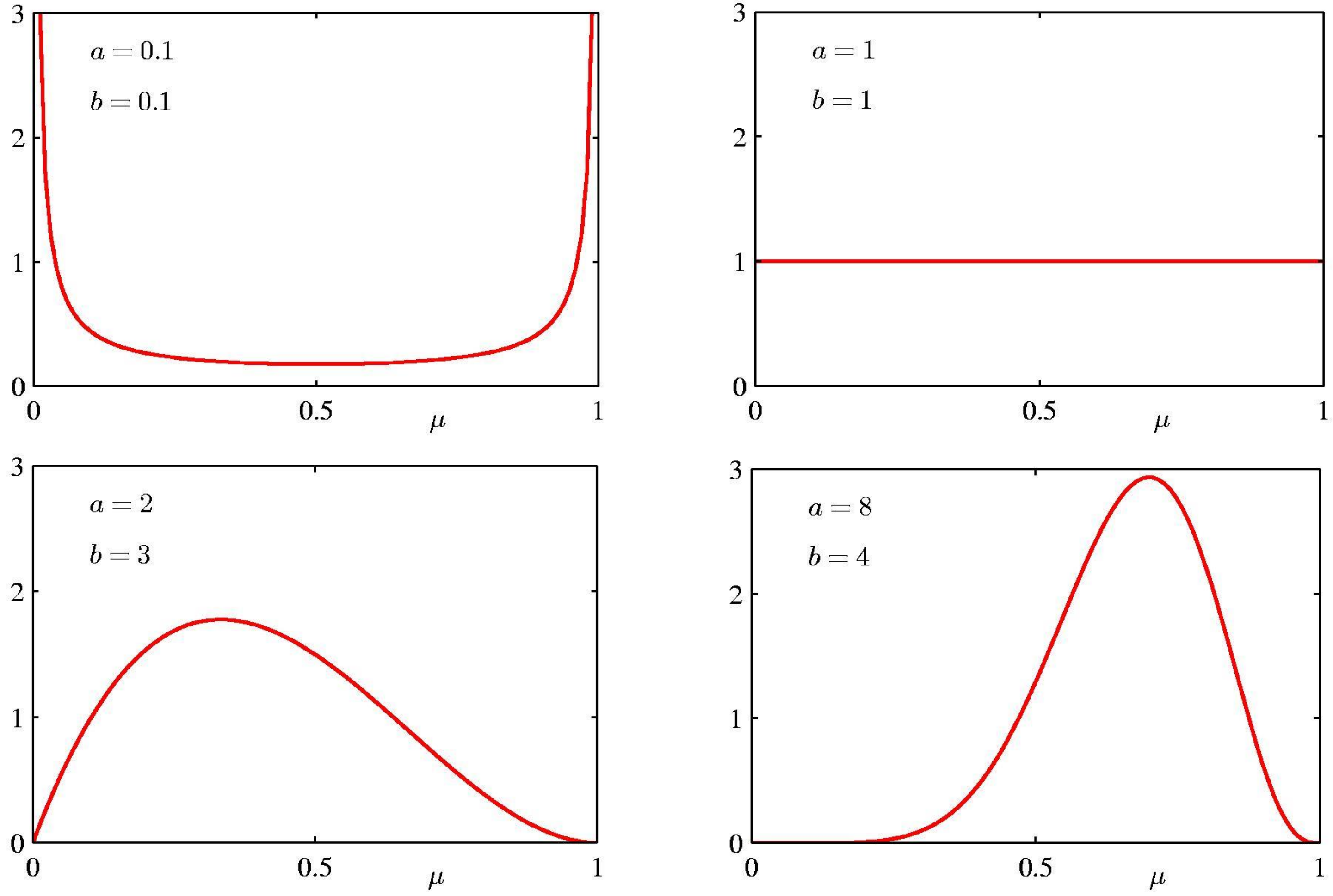
$$= \Gamma(a+b)\int_0^1 t^{a-1}(1-t)^{b-1}dt$$

$$= \Gamma(a+b)B(a,b)$$

$$\Rightarrow \qquad \boxed{B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}$$

Plots of the beta distribution for various values of the hyperparameters a and b.

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$



Main statistics of the beta distribution as a function of the a and b **hyperparameters** (a and b, integers)

$$\mathrm{Beta}(\mu|a,b) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\mu^{a-1}(1-\mu)^{b-1}$$

$$\begin{aligned}
\mathbb{E}(\mu) &= \int_0^1 \mu \ \mathrm{Beta}(\mu|a,b)d\mu \\
&= \int_0^1 \frac{(a+b-1)!}{(a-1)!(b-1)!}\mu^{a}(1-\mu)^{b-1}d\mu \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\mathrm{B}(a+1,b) \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\frac{a!(b-1)!}{(a+b)!} = \frac{a}{a+b}
\end{aligned}$$

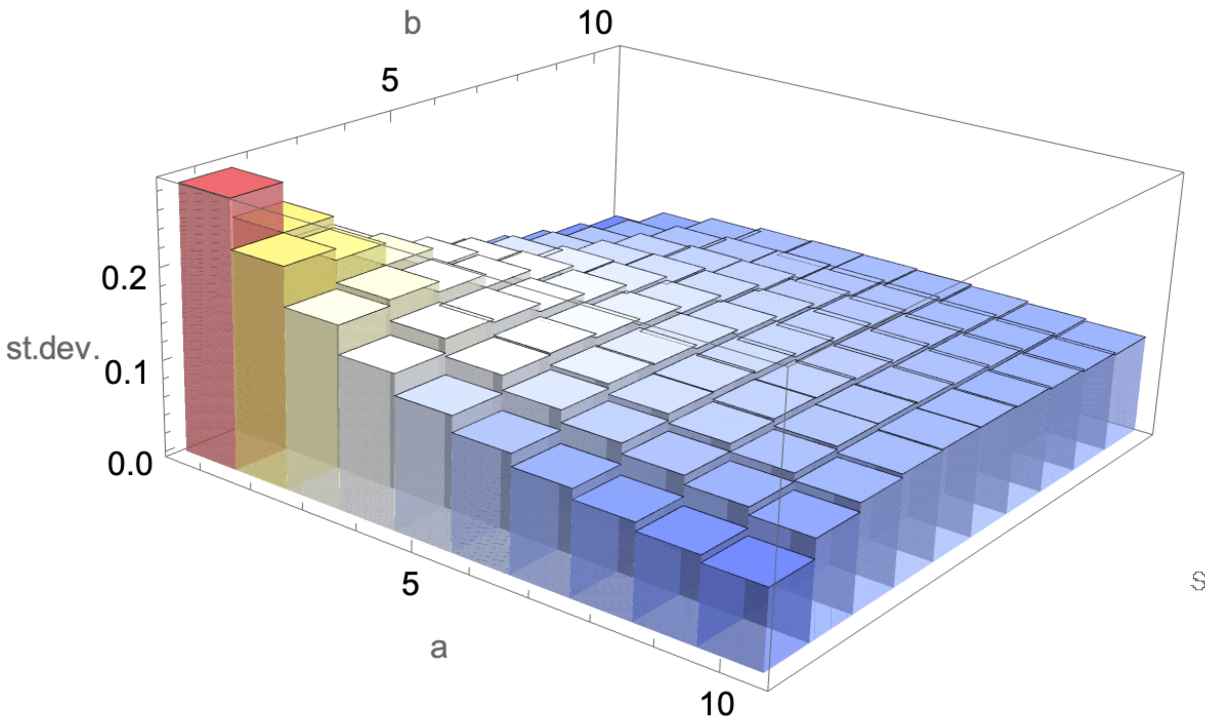
$$\mathrm{var}(\mu) = \mathbb{E}(\mu^2) - \mathbb{E}(\mu)^2 = \frac{ab}{(a+b+1)(a+b)^2}$$

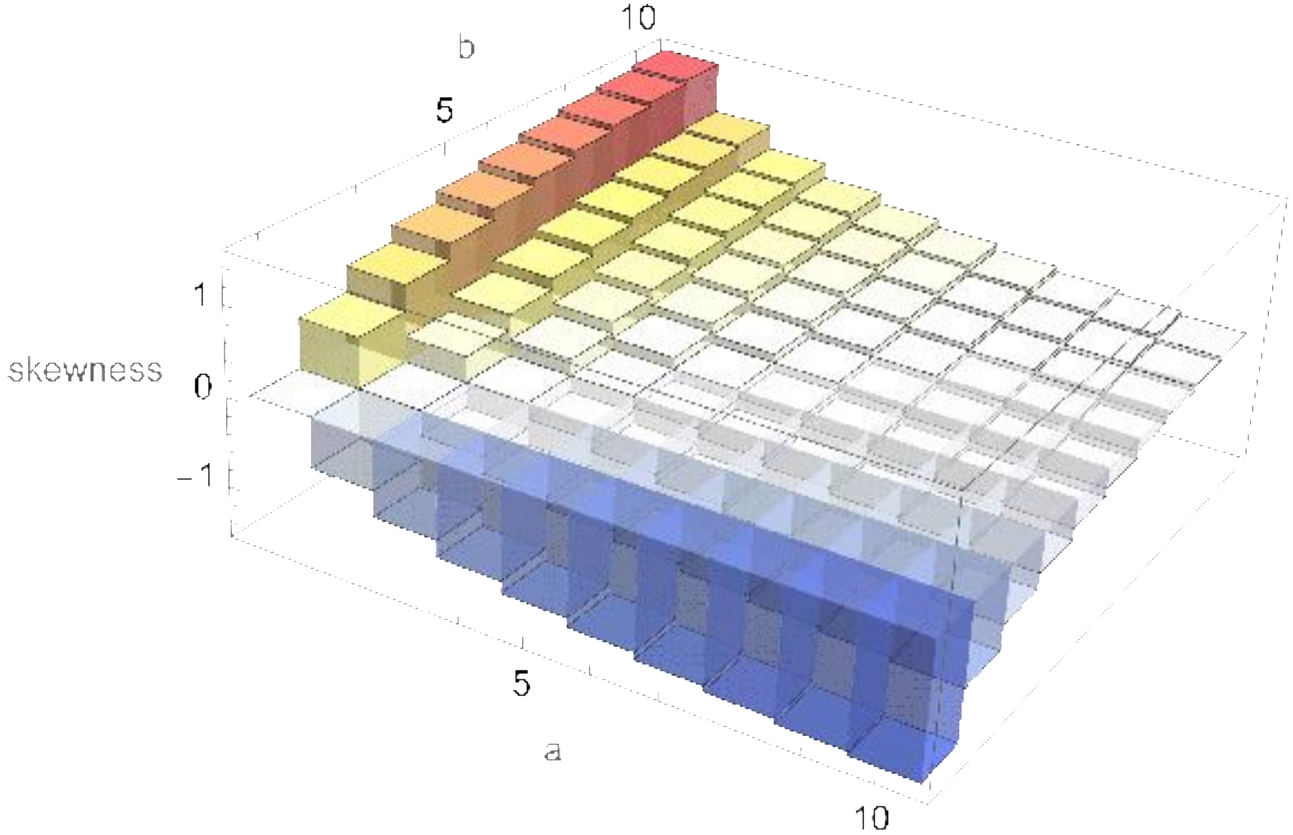
$$\mathrm{skewness}(\mu) = \frac{\mathbb{E}[(\mu-\mathbb{E}(\mu))^3]}{\sigma^2} = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$$

$$\begin{aligned}
\mathbb{E}(\mu^2) &= \int_0^1 \mu^2 \ \mathrm{Beta}(\mu|a,b)d\mu \\
&= \int_0^1 \frac{(a+b-1)!}{(a-1)!(b-1)!}\mu^{a+1}(1-\mu)^{b-1}d\mu \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\mathrm{B}(a+2,b) \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\frac{(a+1)!(b-1)!}{(a+b+1)!} = \frac{a(a+1)}{(a+b+1)(a+b)}
\end{aligned}$$

$$\mathbb{E}(\mu^3) = \frac{a(a+1)}{(a+b+2)(a+b+1)(a+b)}$$

a and *b*, control the values of all moments of the distribution


b
10
5
0.2
st.dev.
0.1
0.0
5
a
10


10
b
5
1
skewness
0
−1
5
a
10

posterior pdf

$$p(\theta|n,N) = \mathrm{Beta}(n+1, N-n+1)$$

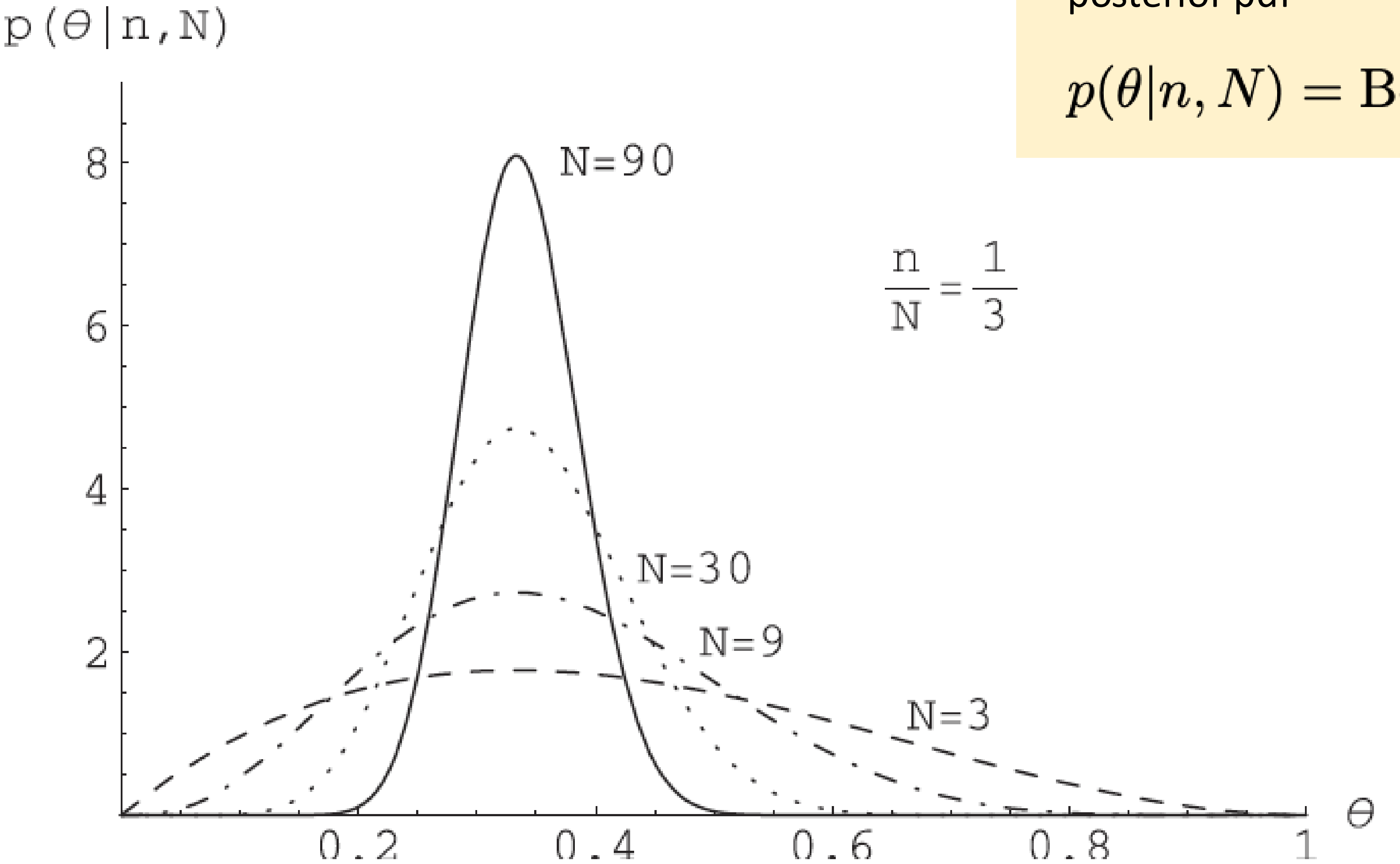


Figure 1. Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

From the knowledge of the posterior pdf we find

$$p(\theta|n, N) = \mathrm{Beta}(n+1, N-n+1)$$

$$\begin{cases} n = a - 1 & \text{number of 1's} \\ N - n = b - 1 & \text{number of 0's} \end{cases}$$

meaning of the hyperparameters

$$\mathbb{E}(\theta) = \frac{n+1}{N+2}$$

biased, asymptotically unbiased, estimator

$$\mathrm{var}(\theta) = \frac{(n+1)(N-n+1)}{(N+3)(N+2)^2}$$

Maximum a posteriori (MAP) estimate – MAP ≠ mean value!

Consider the case with a uniform prior: from the posterior distribution

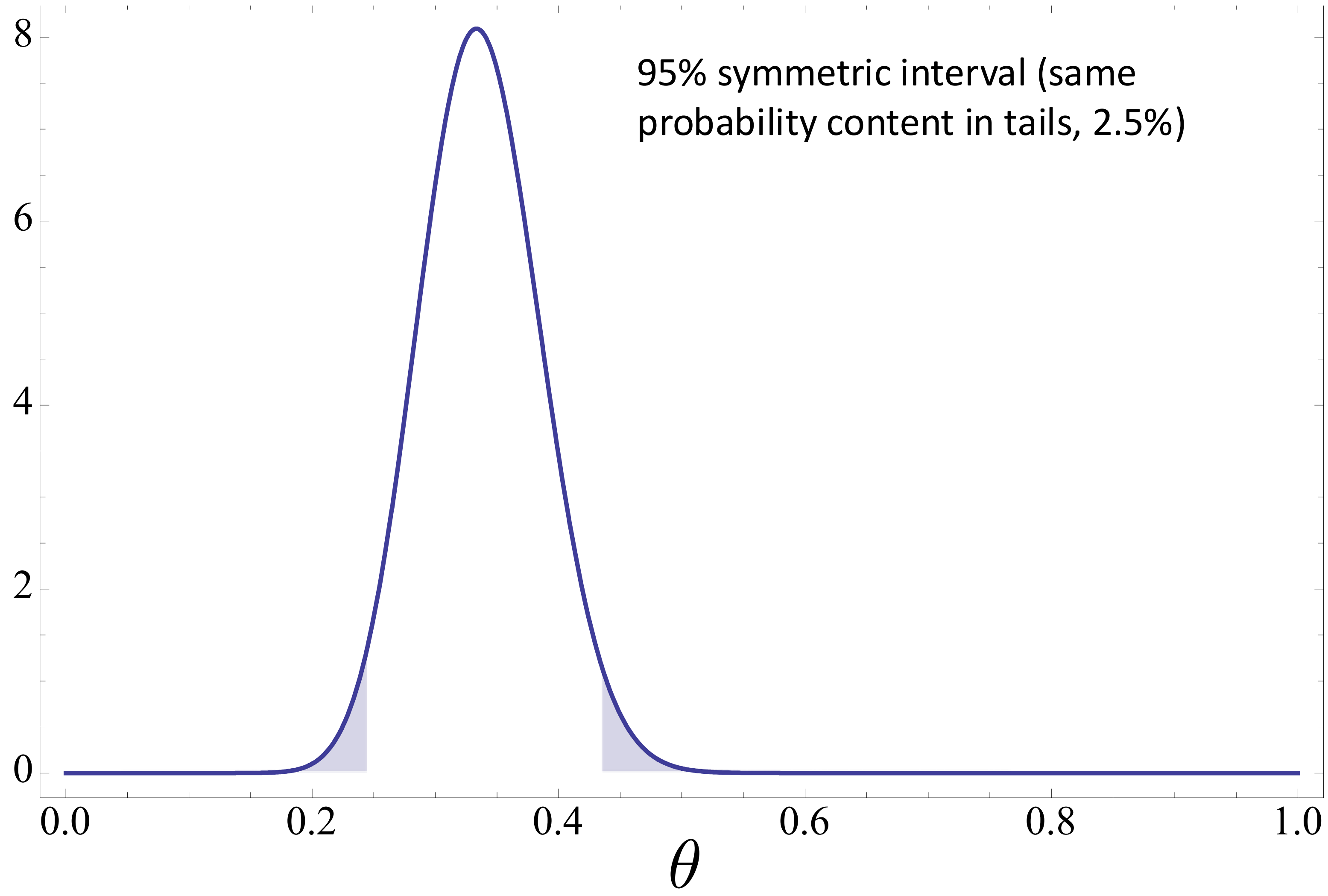
$$p(\theta|n, N) = \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\theta^n$$

we easily find that the posterior pdf is maximized by the parameter value

$$\theta = n/N$$

which is the unbiased estimate of the parameter (unlike the mean value!)

Credible intervals (case of initial uniform prior), the Bayesian analog of confidence intervals.

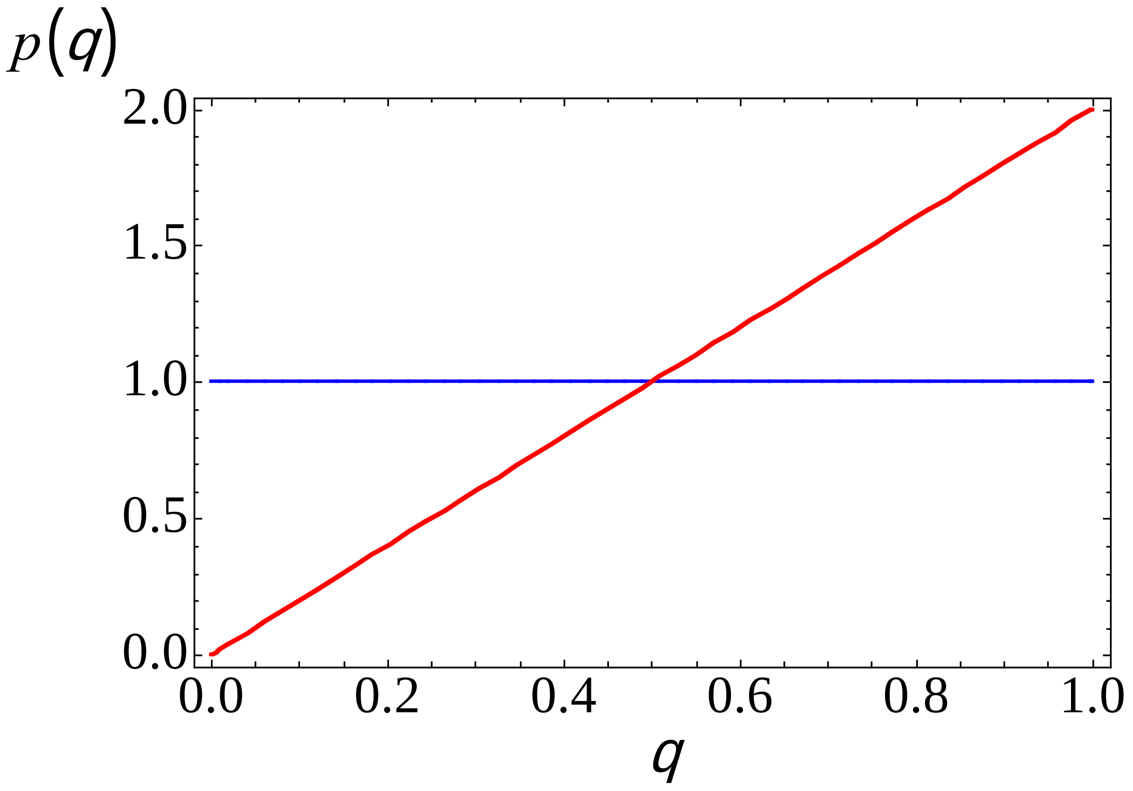


95% symmetric interval (same probability content in tails, 2.5%)

What happens if we try a different prior?

Let’s try with a linear prior

$$p(\theta) = 2\theta$$



$$p(\theta|n,N) = \frac{P(n|\theta,N)}{\int_0^1 P(n|\theta',N)p(\theta')d\theta'}\, p(\theta)$$

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)'^{N-n}\theta'^n\ 2\theta' d\theta'}\, 2\theta = \frac{(1-\theta)^{N-n}\theta^{n+1}}{\int_0^1 (1-\theta)'^{N-n}\theta'^{n+1}\ d\theta'}$$

a beta distribution … again …

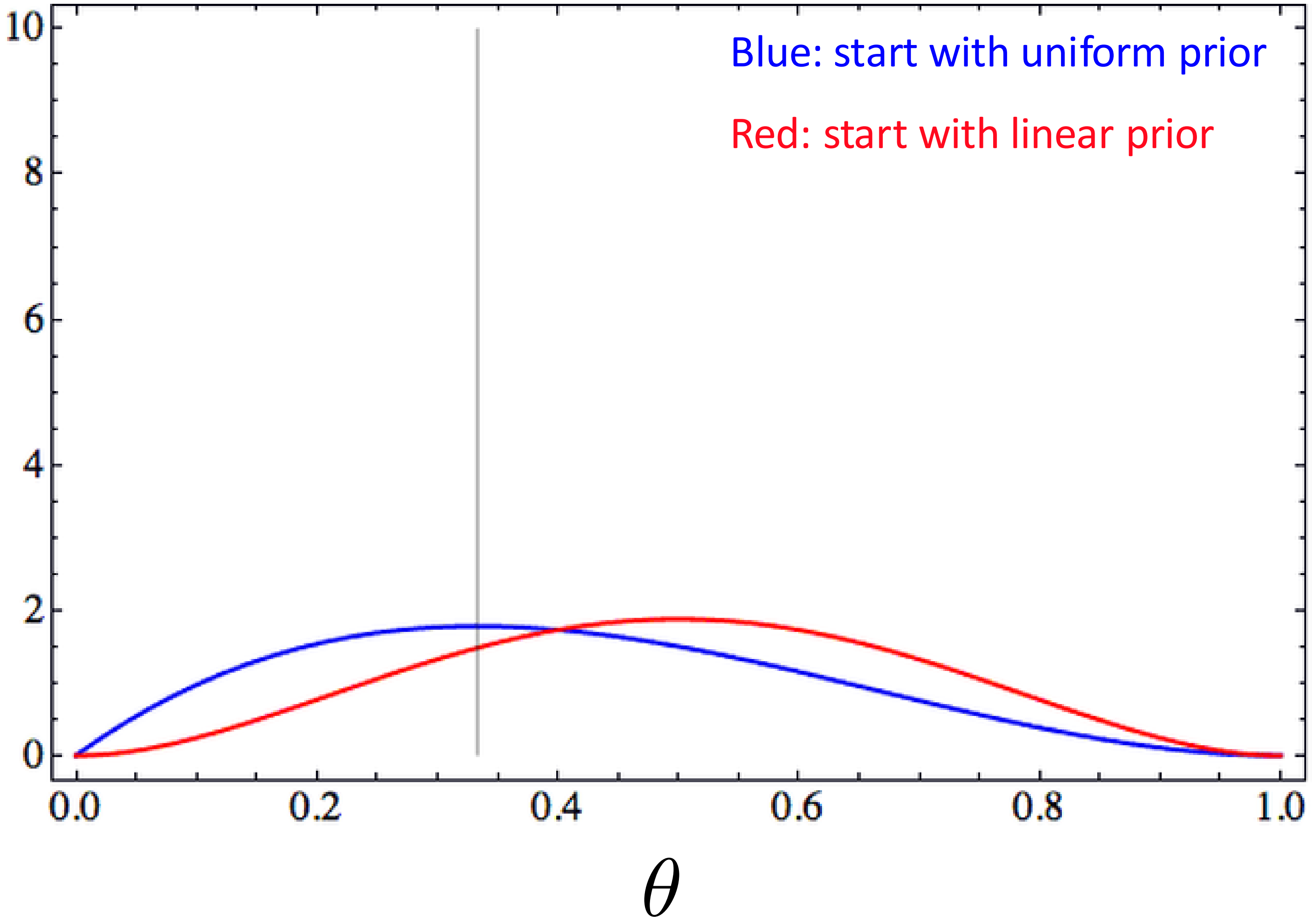
Properties of the new posterior

$$p(\theta|n,N) = \mathrm{Beta}(n+2, N-n+2)$$

$$\begin{cases} n = a-2 & \text{effective number of 1's} \\ N-n = b-1 & \text{effective number of 0's} \end{cases}$$

$$\mathbb{E}(\theta) = \frac{n+2}{N+3}$$

$$\mathrm{var}(\theta) = \frac{(n+2)(N-n+1)}{(N+4)(N+3)^2}$$

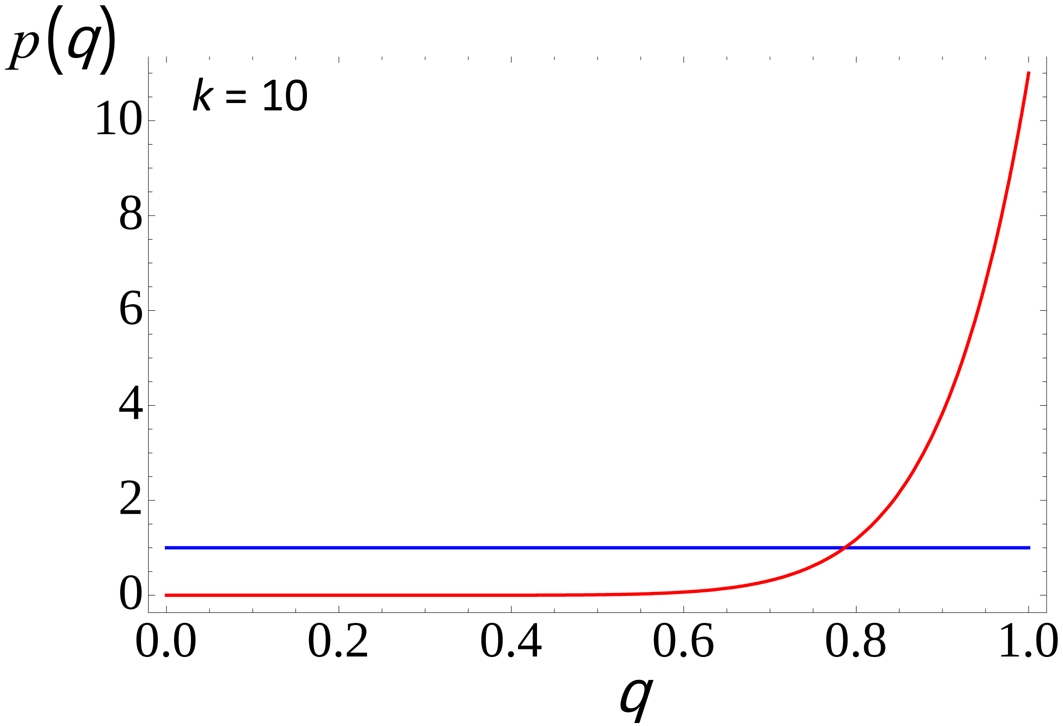


Taking few coin throws, the posterior from the linear prior is considerably biased. The bias disappears when the number of coin throws is large.

Now we try with a very non-uniform prior

We take

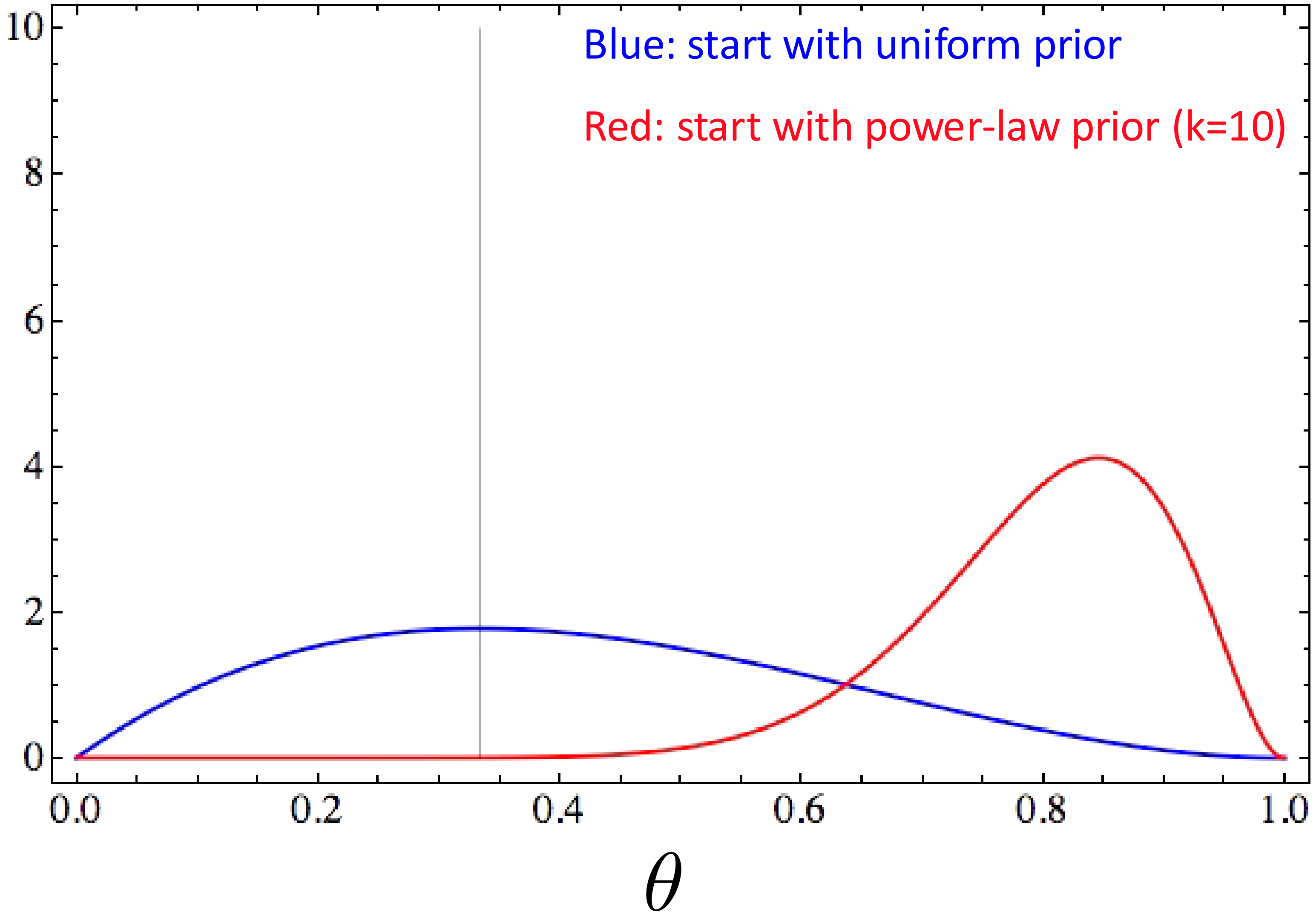
$$p(\theta) = (k+1)\theta^k; \quad k \gg 1$$



$$p(\theta|n,N) = \frac{P(n|\theta,N)}{\int_0^1 P(n|\theta',N)p(\theta')d\theta'}\, p(\theta)$$

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)'^{N-n}\theta'^n\,(k+1)\theta'^k d\theta'}\,(k+1)\theta^k = \frac{(1-\theta)^{N-n}\theta^{n+k}}{\int_0^1 (1-\theta)'^{N-n}\theta'^{n+k}\,d\theta'}$$

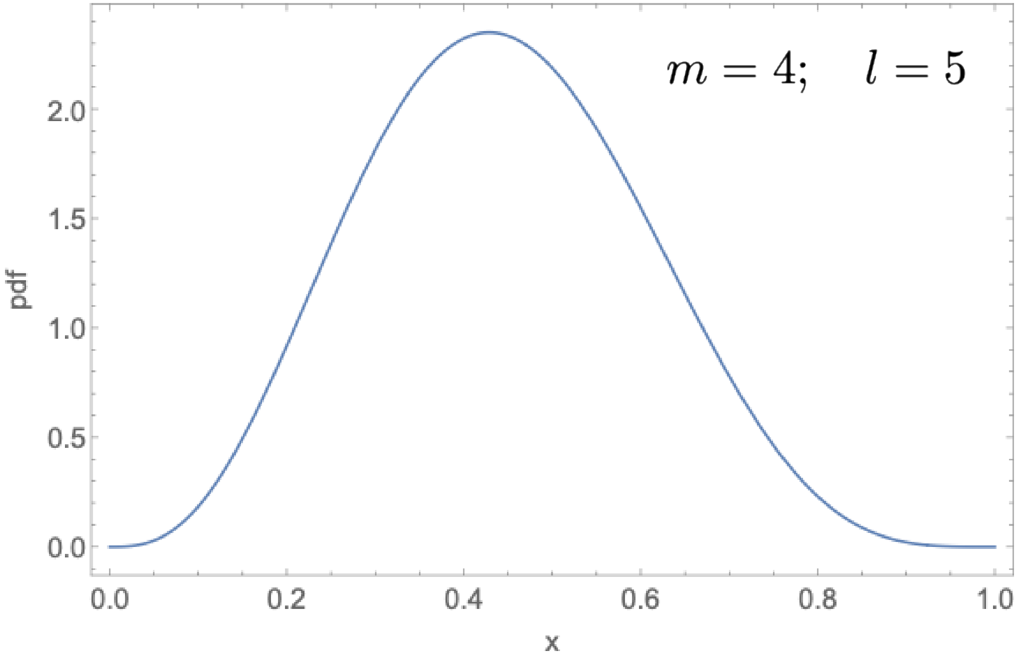
a beta distribution … yet again …



In this case, initial bias due to the prior is very large.

More generally, we can take a Beta prior

$$p(\theta) = \mathrm{Beta}(\theta|m,l) \propto \theta^{m-1}(1-\theta)^{l-1}$$



$$p(\theta|n,N) = \mathrm{Beta}(n+m, N-n+l)$$

$$\begin{cases} n+m & \text{effective number of 1's} \\ N-n+l & \text{effective number of 0's} \end{cases}$$

and

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{n+m}{N+m+l}$$

Lessons learned:

1. The prior information is not neutral, a careful choice of the prior distribution is a necessity.

   Question: how do we choose a prior?

2. If we want to keep all possibilities alive, we must heed the Cromwell’s rule: “Prior probabilities 0 and 1 should be avoided” (Lindley, 1991)

   The reference is to Oliver Cromwell’s phrase:
   I beseech you, in the bowels of Christ, think it possible that you may be mistaken.

3. Convergence as the dataset size grows seems to be granted, however it may be very slow with a bad choice of prior distribution

   Question: is convergence really granted???

Lessons learned - 2:

1. Thanks to the functional shape of the likelihood, a Beta prior leads to a Beta posterior … This property is called *conjugacy*.

2. Thanks to conjugacy, in this case we see that we can feed whole chunks of data to an existing posterior and still find that the sequential learning property of Bayes theorem is satisfied.

3. As the number of observations grows, the posterior is more and more peaked → connection with frequentist statistics

Connection with frequentist statistics:

Consider the following identity involving the (joint and conditional) distributions of parameters and data

$$\mathbb{E}_\theta[\theta] = \int \theta \, p(\theta) d\theta = \int \int \theta \, p(\theta, \mathcal{D}) d\theta d\mathcal{D} = \int \left[\int \theta \, p(\theta|\mathcal{D}) d\theta \right] p(\mathcal{D}) d\mathcal{D}$$

$$= \mathbb{E}_\mathcal{D} \left[\mathbb{E}_\theta [\theta|\mathcal{D}] \right]$$

Prior mean

Posterior mean,
averaged over the
distribution of data

Connection with frequentist statistics - 2:

A similar identity holds when we consider the variance

$$
\begin{aligned}
\mathrm{var}_\theta(\theta) &= \int \theta^2\, p(\theta) d\theta - \left(\mathbb{E}_\theta[\theta]\right)^2 \\
&= \int \left[\int \theta^2\, p(\theta|\mathcal{D}) d\theta \right] p(\mathcal{D}) d\mathcal{D} - \left(\mathbb{E}_\theta[\theta]\right)^2 \\
&= \mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta^2|\mathcal{D}]] - \left(\mathbb{E}_\theta[\theta]\right)^2 \\
&= \mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta^2|\mathcal{D}]] - \mathbb{E}_\mathcal{D}[(\mathbb{E}_\theta[\theta|\mathcal{D}])^2] + \mathbb{E}_\mathcal{D}[(\mathbb{E}_\theta[\theta|\mathcal{D}])^2] - \left(\mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]]\right)^2 \\
&= \mathbb{E}_\mathcal{D}[\mathrm{var}_\theta(\theta|\mathcal{D})] + \mathrm{var}_\mathcal{D}(\mathbb{E}_\theta[\theta|\mathcal{D}])
\end{aligned}
$$

Prior variance

posterior variance, averaged over the distribution of data

variance over data of the posterior mean

$$\mathrm{var}_\theta(\theta) \geq \mathbb{E}_\mathcal{D}[\mathrm{var}_\theta(\theta|\mathcal{D})]$$

the prior variance is greater or equal than the (mean) posterior variance !!!

The Bernstein-Von Mises theorem

- The theorem that grants convergence under very weak hypotheses is the Bernstein-Von Mises theorem. The theorem states that a posterior distribution converges in the limit of infinite data to a multivariate normal distribution centered at the maximum likelihood estimator with covariance matrix given by the normalized Fisher matrix.

- Convergence can only be defined with respect to a frequentist approach (this requires repeated, independent tests of the experimental procedure).

- In the case of nonparametric statistics and for certain probability spaces, the Bernstein-von Mises theorem usually fails.

Predictive power of the posterior distribution

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1,\theta|\mathcal{D})d\theta = \int_0^1 p(x=1|\theta,\mathcal{D})\; p(\theta|\mathcal{D})d\theta =$$

$$= \int_0^1 \theta\; p(\theta|\mathcal{D})d\theta = \mathbb{E}[\theta|\mathcal{D}]$$

$$= \frac{n+m}{N+m+l}$$