Introduction to Bayesian Methods- 2

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

If your experiment needs statistics, you ought to have done a better experiment.

(Ernest Rutherford, as reported by John Hammersley)

Question:

Why do we use statistics in science?

Answer?

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

1. *Example of Bayesian inference*: estimate of the (probability) parameter of the binomial distribution

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

this is the parameter that we want to infer from data

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

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

uniform distribution: the least informative prior

final result is a beta distribution

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

$$B(m,n) = \int_0^1 t^{m-1}(1-t)^{n-1}\, dt$$

$$= \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

beta function

$$p(\theta | n, N) = \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)}(1-\theta)^{N-n}\,\theta^n$$

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

Mathematical digression: relationship between gamma and beta function

$$\Gamma(m)\Gamma(n) = \int_0^\infty s^{m-1} e^{-s}\, ds \int_0^\infty t^{n-1} e^{-t}\, dt$$

$$s = x^2; \qquad t = y^2; \qquad \Rightarrow$$

$$\Gamma(m)\Gamma(n) = 4\int_0^\infty x^{2m-1} e^{-x^2}\, dx \int_0^\infty y^{2n-1} e^{-y^2}\, dy$$

$$x = r\cos\theta; \qquad y = r\sin\theta; \qquad \Rightarrow$$

$$\Gamma(m)\Gamma(n) = 4\int_0^\infty r^{2m+2n-1} e^{-r^2}\, dr \int_0^{\pi/2} \cos^{2m-1}\theta \sin^{2n-1}\theta\, d\theta$$

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

$$= \Gamma(m+n)\int_0^1 t^{m-1}(1-t)^{n-1}\, dt$$

$$= \Gamma(m+n) B(m,n)$$

$$\Rightarrow \qquad B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \qquad \Rightarrow \qquad B(m+1,n+1) = \frac{m!n!}{(m+n+1)!}$$

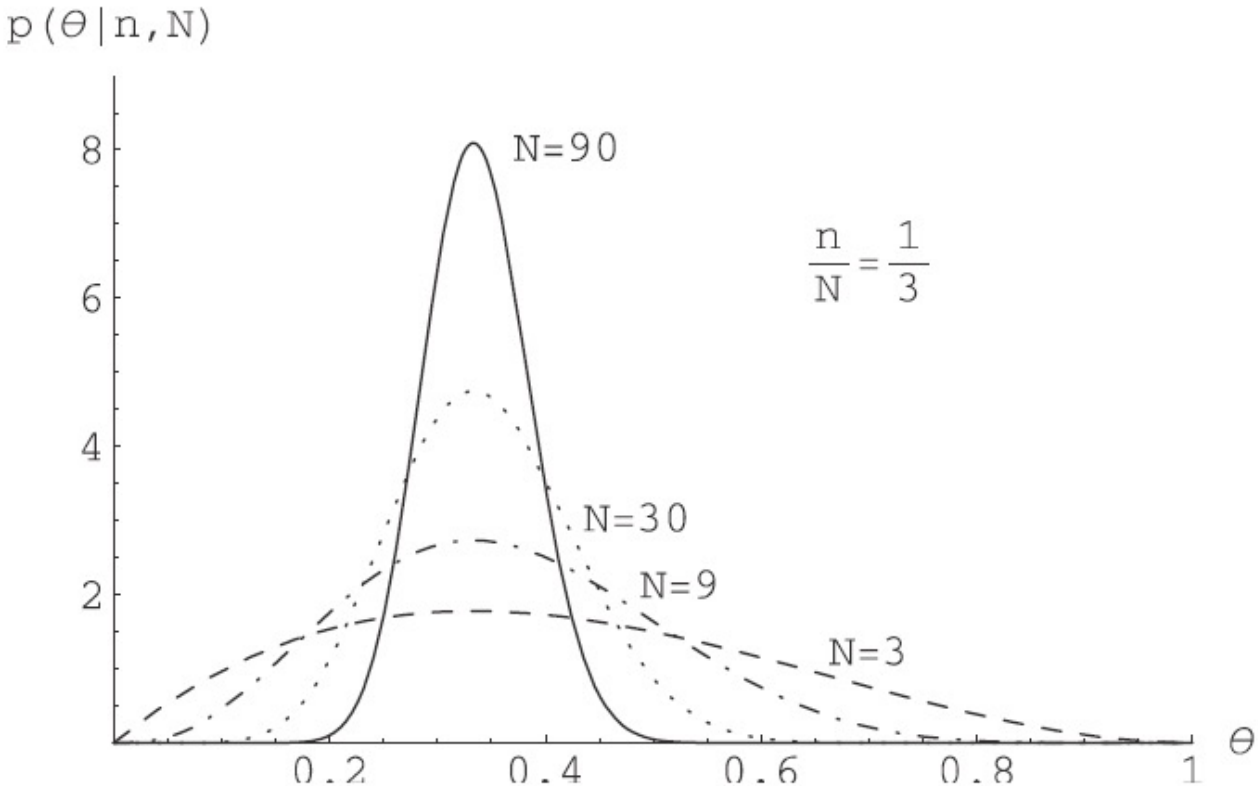


Figure 1. Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

From the knowledge of the posterior pdf we obtain all the momenta of the distribution

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

$$\langle\theta\rangle = \int_0^1 p(\theta \mid n, N)\,\theta d\theta = \frac{(N+1)!}{n!(N-n)!}\int_0^1 (1-\theta)^{N-n}\,\theta^{n+1}\,d\theta$$

$$= \frac{(N+1)!}{n!(N-n)!} B(n+2, N-n+1)$$

$$= \frac{(N+1)!}{n!(N-n)!} \cdot \frac{(n+1)!(N-n)!}{(N+2)!}$$

$$= \frac{n+1}{N+2} \to \frac{n}{N}$$

biased, asymptotically unbiased, estimator

$$\left\langle \theta^2 \right\rangle = \int_0^1 p\left(\theta \mid n, N\right) \theta^2 d\theta = \frac{(N+1)!}{n!(N-n)!} \int_0^1 \left(1-\theta\right)^{N-n} \theta^{n+2}\, d\theta$$

$$= \frac{(N+1)!}{n!(N-n)!} B\left(n+3, N-n+1\right)$$

$$= \frac{(N+1)!}{n!(N-n)!} \cdot \frac{(n+2)!(N-n)!}{(N+3)!}$$

$$= \frac{\left(n+2\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)}$$

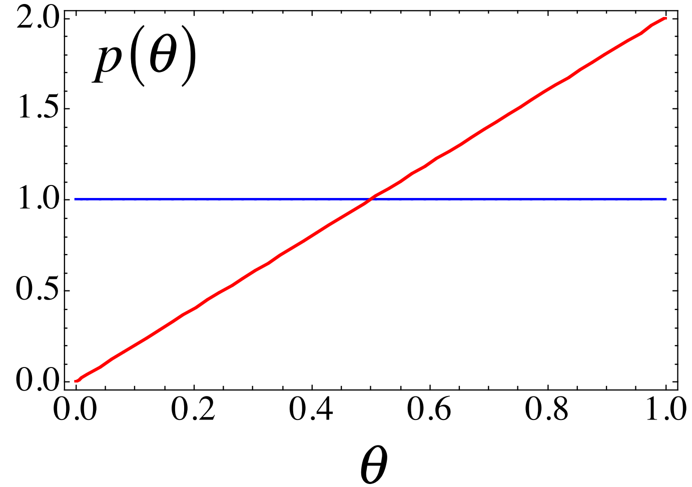
$$\operatorname{var}\theta = \left\langle \theta^2 \right\rangle - \left\langle \theta \right\rangle^2 = \frac{\left(n+2\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)} - \left(\frac{n+1}{N+2}\right)^2 =$$

$$= \frac{\left(N-n+1\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)^3}$$

What happens if we try a different prior?

Let's try with a linear prior

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



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

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

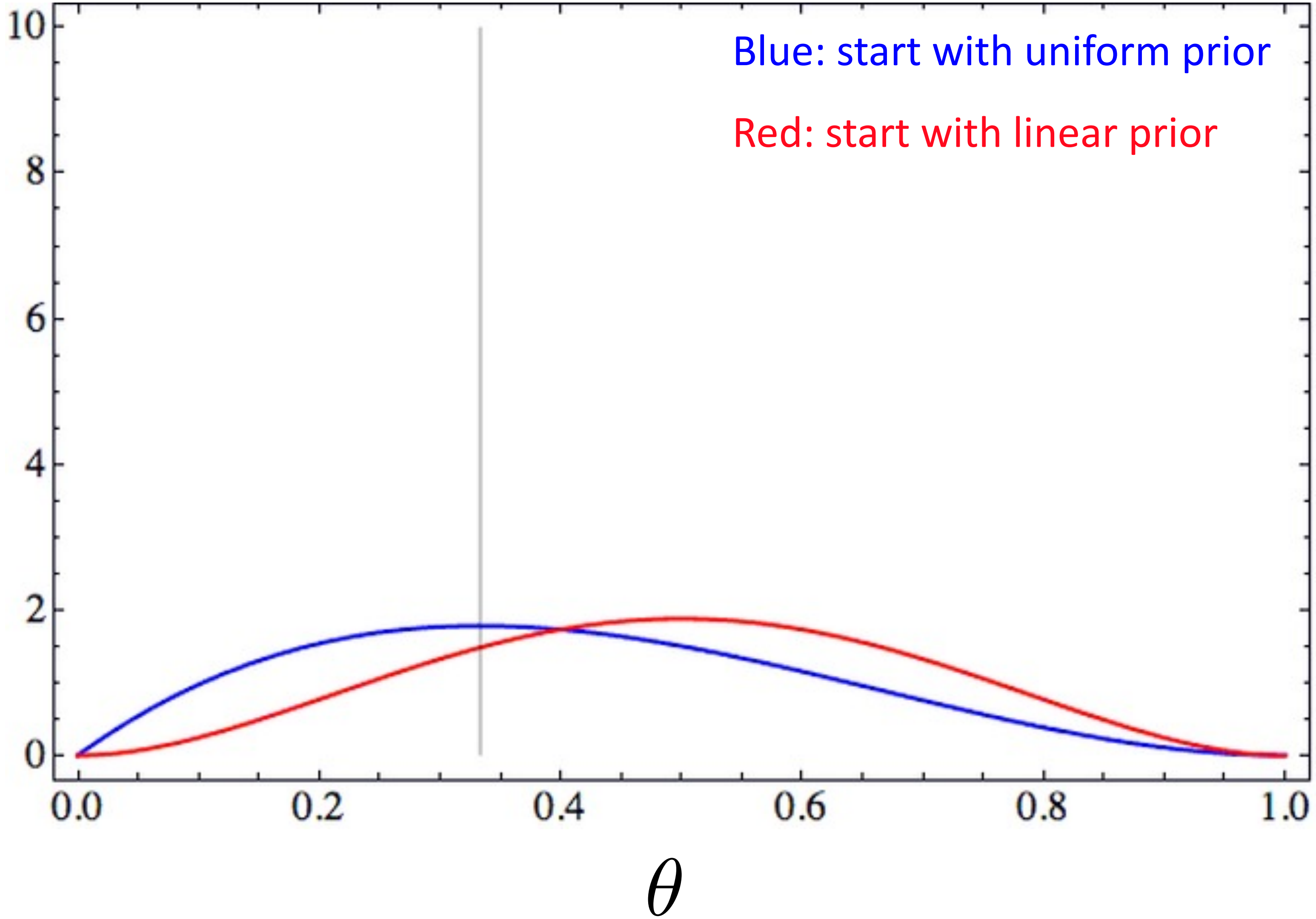
$$p(\theta \mid n, N) = \frac{(N+2)!}{(n+1)!(N-n)!}(1-\theta)^{N-n}\,\theta^{n+1}$$

$$\langle \theta \rangle = \int_0^1 p(\theta \mid n, N)\,\theta d\theta = \frac{(N+2)!}{(n+1)!(N-n)!}\int_0^1 (1-\theta)^{N-n}\,\theta^{n+2}\,d\theta$$

$$= \frac{(N+2)!}{(n+1)!(N-n)!} B(n+3, N-n+1)$$

$$= \frac{(N+2)!}{(n+1)!(N-n)!} \cdot \frac{(n+2)!(N-n)!}{(N+3)!}$$

$$= \frac{n+2}{N+3} \rightarrow \frac{n}{N}$$

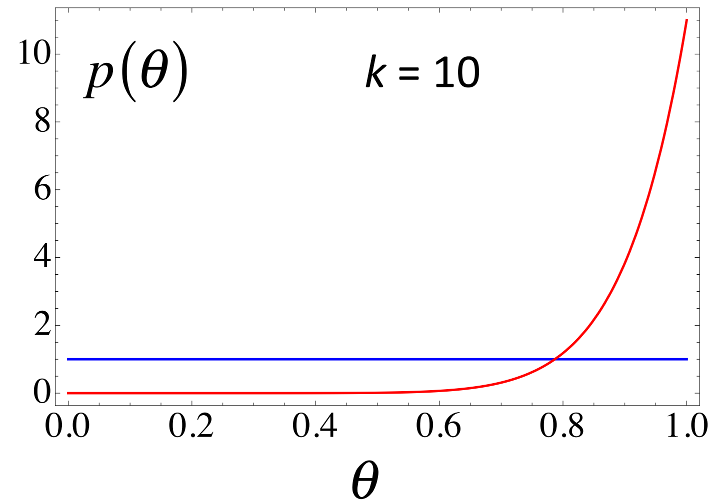


Taking few coin throws, the posterior from the linear prior is considerably biased. The bias disappears when the number of coin throws is large.

Now we try with a very non-uniform prior

We take

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



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

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

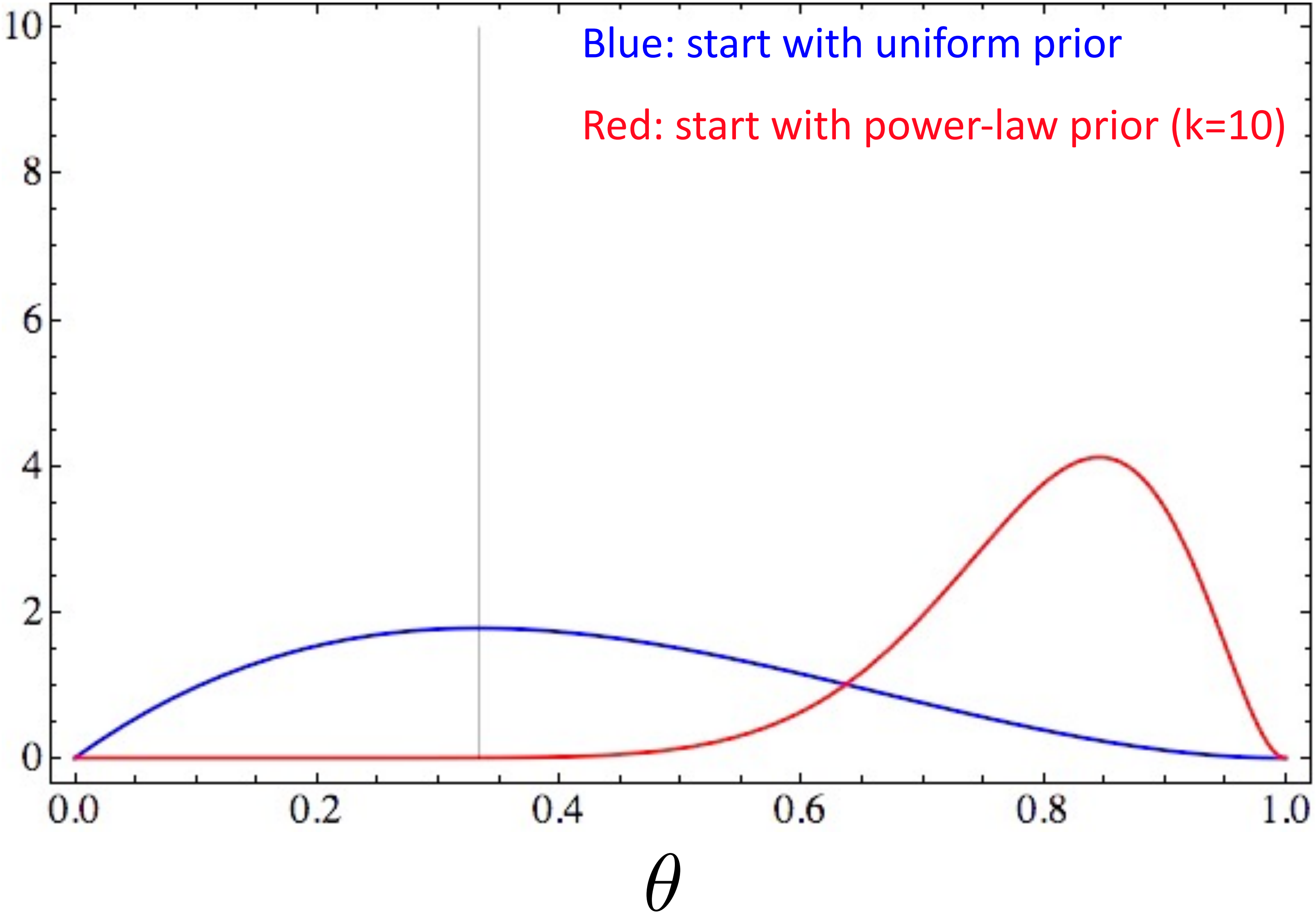
$$p(\theta \mid n,N) = \frac{(N+k+1)!}{(n+k)!(N-n)!}(1-\theta)^{N-n}\,\theta^{n+k}$$

$$\langle\theta\rangle = \int_0^1 p(\theta \mid n,N)\,\theta d\theta = \frac{(N+k+1)!}{(n+k)!(N-n)!}\int_0^1 (1-\theta)^{N-n}\,\theta^{n+k+1}\,d\theta$$

$$= \frac{(N+k+1)!}{(n+k)!(N-n)!}B(n+k+2, N-n+1)$$

$$= \frac{(N+k+1)!}{(n+k)!(N-n)!}\cdot\frac{(n+k+1)!(N-n)!}{(N+k+2)!}$$

$$= \frac{n+k+1}{N+k+2} \to \frac{n}{N}$$



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

Note on posterior distributions:

the relationship between binomial distribution and beta function is quite important and common, and leads to the formal definition of the Beta distribution:

$$\mathrm{B}\left(\theta \middle| a,b\right) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}\left(1-\theta\right)^{b-1}$$

There are other important dualities between distributions. This topic is discussed in depth in

J. M. Bernardo: “Reference Posterior Distributions for Bayesian Inference”, J. R. Statist. Soc. B **41** (1979), 113

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???

The Bernstein-Von Mises theorem

- Convergence can only be defined with respect to a frequentist approach.
- The theorem that grants convergence under very weak hypotheses is the Bernstein-Von Mises theorem.
- It is interesting to note that even here we can find inconsistencies.

Maximum a posteriori (MAP) estimate – MAP is not mean value!

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

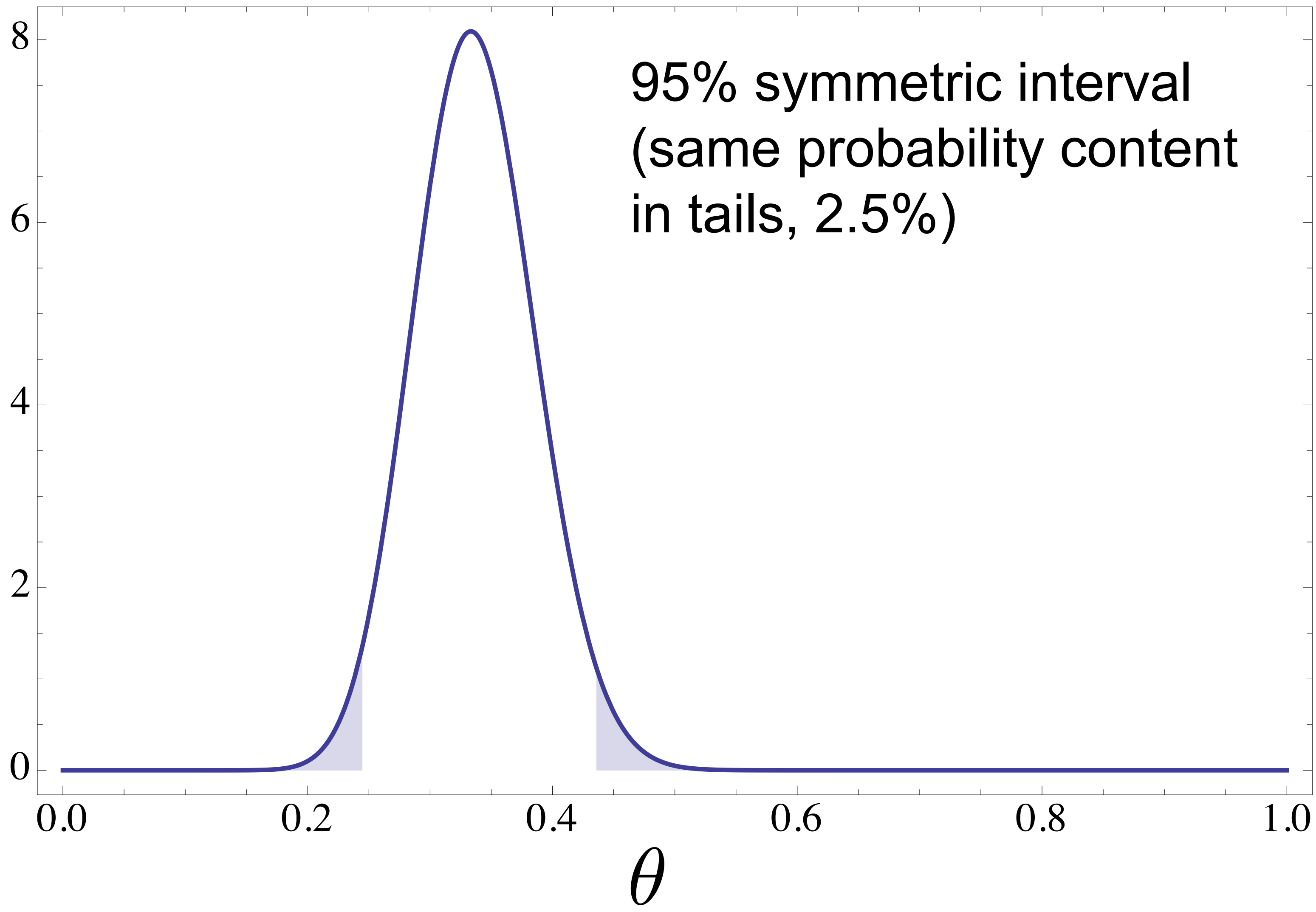
$$p(\theta | n, N) = \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\,\theta^n$$

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), Bayesian analog of confidence intervals.



Example: a decision problem (Skilling 1998)

Let T be the temperature of a liquid which can be either water or ethanol.

1. We suppose first that the liquid is water: then we take a uniform prior distribution for T, between 0 °C and 100 °C
2. The experimental apparatus and the measurement process is defined by the likelihood function **P(D|T,water,I)**. We assume that measurements are uniformly distributed within a range ±5 °C. Therefore **P(D|T,water,I) = 0.1 (°C)**$^{-1}$ in the interval **[T-5°C, T+5°C]**, and zero elsewhere.
3. We take a single measurement **D = -3°C**.

4. The evidence *p*(*D*) is*

$$p(D|\text{water}, I) = \int_T p(D|T, \text{water}, I)p(T)dT$$
$$= \int_{0^\circ\text{C}}^{2^\circ\text{C}} \frac{(^\circ\text{C})^{-1}}{10} \frac{(^\circ\text{C})^{-1}}{100} dT(^\circ\text{C}) = 0.002(^\circ\text{C})^{-1}$$

5. Using Bayes' theorem we find

$$p(T|D, \text{water}, I) = \frac{p(D|T, \text{water}, I)}{p(D, \text{water}, I)} p(T|\text{water}, I) = \frac{0.1(^\circ\text{C})^{-1}}{0.002(^\circ\text{C})^{-1}} 0.01(^\circ\text{C})^{-1}$$
$$= 0.5(^\circ\text{C})^{-1} \qquad (0^\circ\text{C} < T < 2^\circ\text{C})$$

* notice that in this case the likelihood is a pdf: the reason is that *D* is a continuous variable

Now suppose that the liquid is ethanol, so that the temperature range is -80°C<T<80°C

1. $p(T) = (160°C)^{-1}$ in -80°C < T < 80°C.
2. $p(D|T,\text{ethanol},I) = 0.1\ (°C)^{-1}$ in [T-5°C, T+5°C], and zero elsewhere.
3. We take a single measurement D = -3°C.
4. The evidence $p(D,\text{ethanol},I)$ is

$$p(D|\text{ethanol}, I) = \int_T p(D|T, \text{ethanol}, I) p(T|\text{ethanol}, I) dT = \int_{-8^\circ\text{C}}^{2^\circ\text{C}} \frac{(^\circ\text{C})^{-1}}{10} \frac{(^\circ\text{C})^{-1}}{160} dT (^\circ\text{C}) = 0.00625 (^\circ\text{C})^{-1}$$

5. Using Bayes' theorem we find

$$p(T|D, \text{ethanol}, I) = \frac{p(D|T, \text{ethanol}, I)}{p(D, \text{ethanol}, I)} p(T|\text{ethanol}, I) = \frac{0.1 (^\circ\text{C})^{-1}}{0.00625 (^\circ\text{C})^{-1}} \frac{1}{160} (^\circ\text{C})^{-1}$$
$$= 0.1 (^\circ\text{C})^{-1} \qquad (-8^\circ\text{C} < T < 2^\circ\text{C})$$

Assuming a prior for the water-ethanol choice, we can discriminate between water and ethanol

$$P_{water} = P_{ethanol} = 0.5$$

With this prior assumption we find,

$$\begin{aligned} P(\mathrm{water}|D,I) &= \frac{p(D|\mathrm{water},I)}{p(D|\mathrm{water},I)P(\mathrm{water}|I) + p(D|\mathrm{ethanol},I)P(\mathrm{ethanol}|I)} P(\mathrm{water}|I) \\ &= \frac{p(D|\mathrm{water},I)}{p(D|\mathrm{water},I) + p(D|\mathrm{ethanol},I)} \end{aligned}$$

and therefore the ratio of the posteriors is given by the Bayes' factor

$$\frac{P(\mathrm{water}|D,I)}{P(\mathrm{ethanol}|D,I)} = \frac{p(D|\mathrm{water},I)}{p(D|\mathrm{ethanol},I)}$$

We have found earlier that

$$p(D|\text{water}, I) = 0.002(^\circ C))^{-1}$$
$$p(D|\text{ethanol}, I) = 0.00625(^\circ C))^{-1}$$

therefore the Bayes factor is

$$B = \frac{P(\text{water}|D, I)}{P(\text{ethanol}|D, I)} = \frac{p(D|\text{water}, I)}{p(D|\text{ethanol}, I)} = 3.125$$

and we conclude that the observation favors the hypothesis of liquid ethanol.

$\log_{10}(B)$	B	Evidence support
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Interpretation of the Bayes factor *B* as evidence support according to Jeffreys (1961), in half units on a scale of $\log_{10}$.

In the case of the water-ethanol problem, and according to Jeffreys' categories, the preference for ethanol is "not worth more than a bare mention", although it happens to be in the upper part of the range.

In 1995, Kass and Raftery noted that *it can be useful to consider twice the natural logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics* and therefore proposed a different interpretation

$2\log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

$$B_{10} = \frac{P(D|H_1)}{P(D|H_0)}$$

Here 1 denotes the alternative hypothesis and 0 the null hypothesis

Example of Bayesian parameter estimation: analytical straight-line fit

$$y_i = ax_i + b + \varepsilon_i$$

y_i measured value

x_i independent variable ("exactly" known)

a, b fit parametes: eventually we expect to find pdf's for these parameters

ε_i statistical error

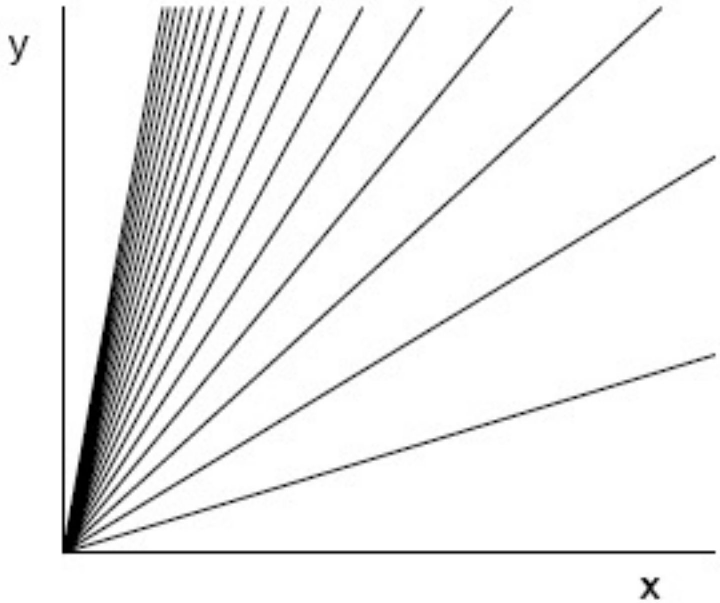
$$\langle \varepsilon_i \rangle = 0; \quad \langle \varepsilon_i^2 \rangle = \sigma^2 \quad \Rightarrow$$

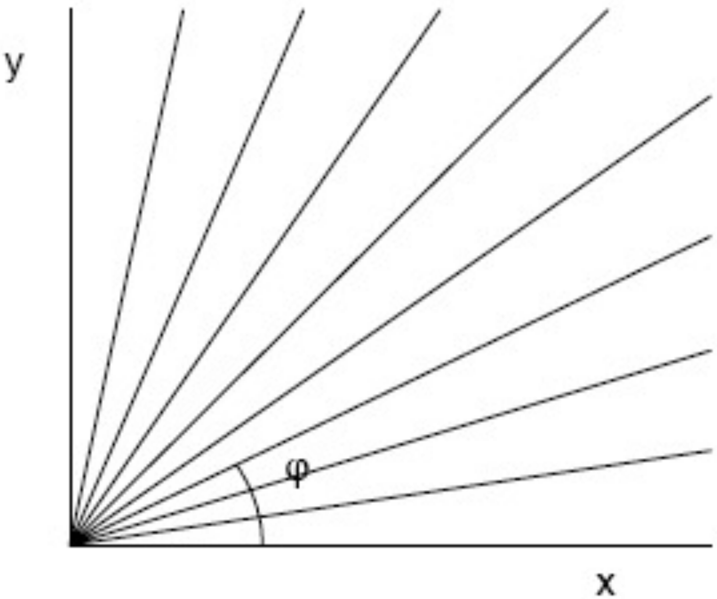
the statistical measurement error has a Gaussian distribution

setting up the likelihood

$$p\left(\mathbf{y} \mid a,b,\mathbf{x},\sigma\right) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]$$

prior angular distribution





uniform *a* uniform angle

The uniform distribution of a introduces an angular bias. The least informative choice corresponds to a uniform angular distribution

$$p_\varphi(\varphi) = \frac{1}{\pi}; \quad -\frac{\pi}{2} \le \varphi < \frac{\pi}{2}$$

and we obtain the distribution of *a* with the transformation method:

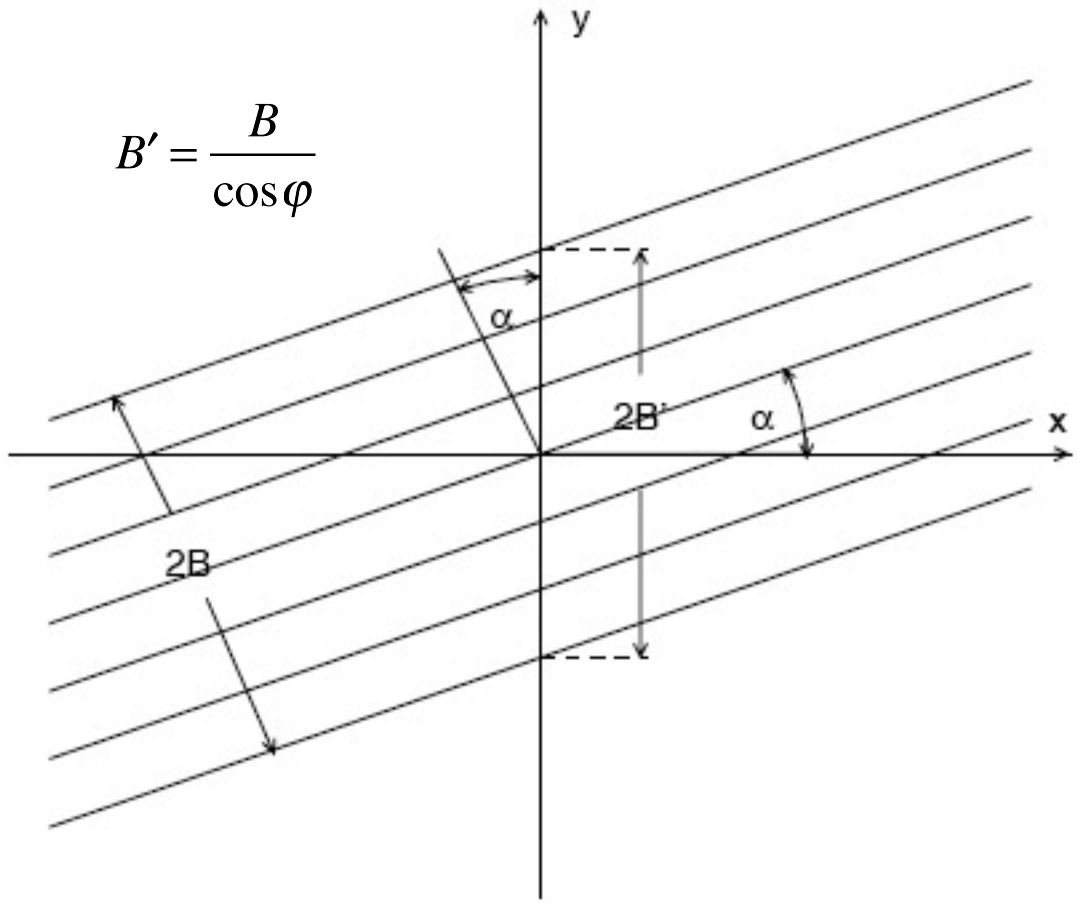
$$a = \tan\varphi$$

$$\Rightarrow \quad p_\varphi(\varphi)d\varphi = p_a(a)da = p_a(a)d(\tan\varphi) = p_a(a)\sec^2\varphi d\varphi$$

$$\Rightarrow \quad p_a(a) = \frac{1}{\pi\sec^2\varphi} = \frac{1}{\pi(1+\tan^2\varphi)} = \frac{1}{\pi(1+a^2)}$$

prior distribution of *b*: improper uniform distribution, related to the distribution of *a*

$$B' = \frac{B}{\cos\varphi}$$



$$p(b \mid a=0) = \frac{1}{2B}; \quad p(b \mid a) = \frac{1}{2B'} = \frac{\cos\varphi}{2B} = \frac{1}{2B} \cdot \frac{1}{\sqrt{1+a^2}}$$

we obtain the posterior from Bayes' theorem

$$p(a,b \mid \mathbf{y},\mathbf{x},\boldsymbol{\sigma}) = \frac{p(\mathbf{y} \mid a,b,\mathbf{x},\boldsymbol{\sigma})}{\int_{-\infty}^{+\infty} da \int_{-B/\cos\varphi}^{B/\cos\varphi} db \; p(\mathbf{y} \mid a,b,\mathbf{x},\boldsymbol{\sigma}) \cdot p(a,b)} \cdot p(a,b)$$

where the prior is

$$p(a,b) = p(b \mid a) \cdot p(a) = \left(\frac{1}{2B} \cdot \frac{1}{\sqrt{1+a^2}} \right) \left(\frac{1}{\pi\left(1+a^2\right)} \right)$$

$$\propto \frac{1}{\left(1+a^2\right)^{3/2}}$$

finally we find

$$p(a,b\,|\,\mathbf{y},\mathbf{x},\sigma)=\frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-ax_i-b)^2\right]}{\left\{\int_{-\infty}^{+\infty}da\int_{-B/\cos\varphi}^{B/\cos\varphi}db\,\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-ax_i-b)^2\right]\cdot\frac{1}{(1+a^2)^{3/2}}\right\}}\cdot\frac{1}{(1+a^2)^{3/2}}$$

$$\approx\frac{\frac{1}{(1+a^2)^{3/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-ax_i-b)^2\right]}{\left\{\int_{-\infty}^{+\infty}\frac{da}{(1+a^2)^{3/2}}\int_{-\infty}^{+\infty}db\,\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-ax_i-b)^2\right]\right\}}$$

This expression has a partly Gaussian structure, and we shall rearrange the quadratic expression in the exponential.

To Be Continued ...