Introduction to Bayesian Statistics - 2

PhD Physics course (XXVIII ciclo)

Università di Trieste

Edoardo Milotti

Sunrise and eclipse over Graz on Jan. 4th 2011 (Robert Pölzl)

What is the probability that the sun rises tomorrow? (Laplace)

$$0 \leq \theta \leq 1$$

is the probability that the sun rises tomorrow, and we assume this probability to be uniformly distributed between 0 and 1.

Then

Probability of the event "the sun rises tomorrow" and that hypothesis H_k is true

$$P(S) = \sum_k P(SH_k)$$

Probability of the event "the sun rises tomorrow"

$$= \sum_k P(S|H_k)P(H_k)$$

$$\rightarrow \int_0^1 P(S|\theta)p(\theta)d\theta$$

hypothesis H_k corresponds to a certain value of p

$$P(S|\theta,N)=\theta$$

Probability that "the sun rises tomorrow", if it already did so N times

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

$$=\frac{\theta^N}{\int_0^1 \theta^N d\theta}=(N+1)\theta^N$$

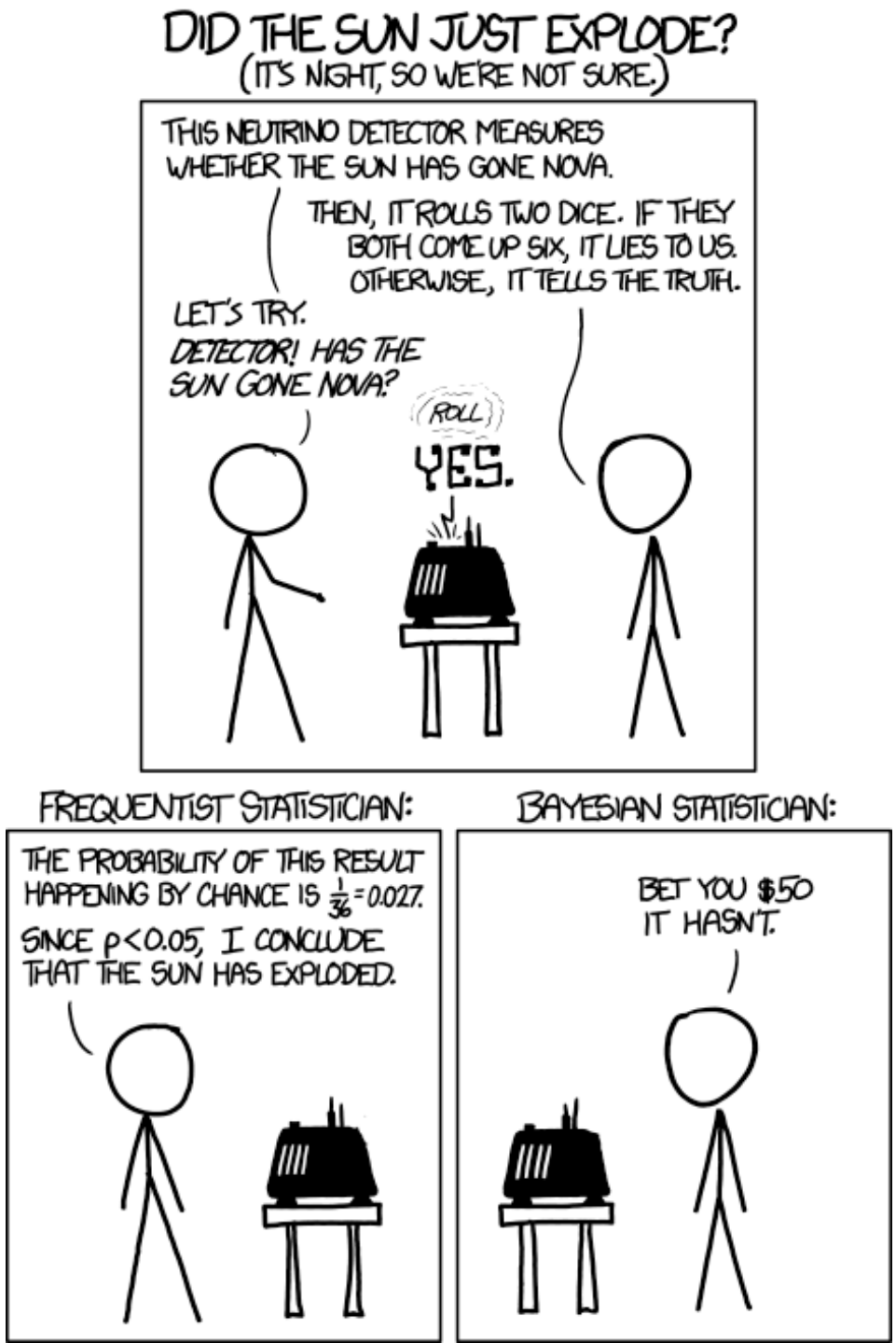
$$P(S|N)=\int_0^1 P(S|\theta,N)p(\theta|N)d\theta=\int_0^1 \theta\cdot(N+1)\theta^N d\theta=\frac{N+1}{N+2}$$

$$P(S|N) = \frac{N+1}{N+2} = 1 - \frac{1}{N+2}$$

Then, for a 27-year old person (about 10000 days), having observed that the sun rises every day means that

$$P(S|N) \approx 1 - \frac{1}{10000} \approx 0.99999$$

or also, that the probability that the sun does not rise tomorrow is about 1/10000 = 0.0001



from http://xkcd.com

Bayesian inference

$$P\left(A_k \mid B\right) = \frac{P\left(B \mid A_k\right) \cdot P\left(A_k\right)}{\sum_{k=1}^{N} P\left(B \mid A_k\right) \cdot P\left(A_k\right)}$$

$$= \frac{P\left(B \mid A_k\right)}{\sum_{k=1}^{N} P\left(B \mid A_k\right) \cdot P\left(A_k\right)} \cdot P\left(A_k\right)$$

$$P\left(H_k \mid D, I\right) = \frac{P\left(D \mid H_k, I\right)}{\sum_{k=1}^{N} P\left(D \mid H_k, I\right) \cdot P\left(H_k \mid I\right)} \cdot P\left(H_k \mid I\right)$$

$$P\left(H_k \mid D, I\right) = \frac{P\left(D \mid H_k, I\right)}{\sum_{k=1}^{N} P\left(D \mid H_k, I\right) \cdot P\left(H_k \mid I\right)} \cdot P\left(H_k \mid I\right)$$

(Posterior probability that *k*-th hypothesis is true, when we observe data D, with prior information I)

=

(Probability of observing data D, given the *k*-th hypothesis) / Normalization

.

(Prior probability that *k*-th hypothesis is true)

$$P(H_k \mid D,I) = \frac{P(D \mid H_k,I)}{P(D \mid I)} \cdot P(H_k \mid I)$$

$$= \frac{P(D \mid H_k,I)}{\sum_{k=1}^{N} P(D \mid H_k,I) \cdot P(H_k \mid I)} \cdot P(H_k \mid I)$$

prior probability $P(H_k,I)$

posterior probability $P(H_k \mid D,I)$

likelihood $P(D \mid H_k,I)$

evidence (normalizing factor) $P(D \mid I) = \sum_{k=1}^{N} P(D \mid H_k,I) \cdot P(H_k \mid I)$

Testing hypotheses

$$P(H_k \mid D, I) = \frac{P(D \mid H_k, I)}{P(D \mid I)} \cdot P(H_k \mid I)$$

likelihood ratio

$$\frac{P(H_k \mid D, I)}{P(H_n \mid D, I)} = \left(\frac{P(D \mid H_k, I)}{P(D \mid H_n, I)} \right) \cdot \left(\frac{P(H_k \mid I)}{P(H_n \mid I)} \right)$$

Bayes' factor

When prior probabilities are the same (equally probable hypotheses), the posterior probability ratio depends only on the Bayes' factor:

$$\frac{P(H_k \mid D, I)}{P(H_n \mid D, I)} = \left(\frac{P(D \mid H_k, I)}{P(D \mid H_n, I)} \right)$$

Recursivity in a parameter estimate

$$P_2\left(\theta \mid \{d_1,d_2\},I\right) = \frac{P\left(\{d_1,d_2\} \mid \theta,I\right)}{P\left(\{d_1,d_2\} \mid I\right)} \cdot P_0\left(\theta \mid I\right)$$

$$P\left(\{d_1,d_2\} \mid \theta,I\right) = P\left(d_2 \mid d_1,\theta,I\right) \cdot P\left(d_1 \mid \theta,I\right)$$

$$= P\left(d_2 \mid \theta,I\right) \cdot P\left(d_1 \mid \theta,I\right)$$

here we assume that successive data are independent

the same trick works for the evidence

$$P\left(\{d_1,d_2\} \mid I\right) = P\left(d_2,I\right) \cdot P\left(d_1 \mid I\right)$$

$$P_2\left(\theta \mid \{d_1,d_2\},I\right) = \left(\frac{P\left(d_2 \mid \theta,I\right)}{P\left(d_2,I\right)}\right) \cdot \left(\frac{P\left(d_1 \mid \theta,I\right)}{P\left(d_1 \mid I\right)} \cdot P_0\left(\theta \mid I\right)\right)$$

$$= \frac{P\left(d_2 \mid \theta,I\right)}{P\left(d_2,I\right)} \cdot P_1\left(\theta \mid I\right)$$

A simple exercise (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°. 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

$$P\left(D\middle|water,I\right)=\int_{T}P\left(D\middle|T,water,I\right)P\left(T\right)dT$$

$$=\int_{0°C}^{2°C}\frac{\left(°C\right)^{-1}}{10}\cdot\frac{\left(°C\right)^{-1}}{100}dT\left(°C\right)=0.002\left(°C\right)^{-1}$$

5. Using Bayes' theorem we find

$$P\left(T\middle|D,water,I\right)=\frac{P\left(D\middle|T,water,I\right)}{P\left(D,water,I\right)}P\left(T\right)=\frac{0.1\left(°C\right)^{-1}}{0.002\left(°C\right)^{-1}}0.01\left(°C\right)^{-1}$$

$$=0.5\left(°C\right)^{-1}\qquad\left(0°C<T<2°C\right)$$

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,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,ethanol,I) is

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

5. Using Bayes' theorem we find

$$P(T|D,ethanol,I)=\frac{P(D|T,ethanol,I)}{P(D,ethanol,I)}P(T)=\frac{0.1(°C)^{-1}}{0.00625(°C)^{-1}}\frac{1}{160}(°C)^{-1}=0.1(°C)^{-1}$$

$$(-8°C<T<2°C)$$

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

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

Indeed,

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

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

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

We have found earlier that

$$P\left(D\middle|water\right)=0.002\left(^{\circ}C\right)^{-1}$$
$$P\left(D\middle|ethanol\right)=0.00625\left(^{\circ}C\right)^{-1}$$

therefore

$$\frac{P\left(ethanol\middle|D,I\right)}{P\left(water\middle|D,I\right)}=\frac{P\left(D\middle|ethanol,I\right)}{P\left(D\middle|water,I\right)}=3.125$$

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.

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.

From discrete sets of hypothesis to the continuum. The Bayes' theorem in the context of parameter estimation.

$$P(H_k \mid D,I) = \frac{P(D \mid H_k,I)}{P(D \mid I)} \cdot P(H_k \mid I) = \frac{P(D \mid H_k,I)}{\sum_{k=1}^{N} P(D \mid H_k,I) \cdot P(H_k \mid I)} \cdot P(H_k \mid I)$$

$$dP(\theta \mid D,I) = \frac{P(D \mid \theta,I)}{\int_{\Theta} P(D \mid \theta,I) \cdot dP(\theta \mid I)} \cdot dP(\theta \mid I)$$

$$\frac{dP(\theta \mid D,I)}{d\theta} = \frac{P(D \mid \theta,I)}{\int_{\Theta} P(D \mid \theta,I) \cdot \frac{dP(\theta \mid I)}{d\theta} d\theta} \cdot \frac{dP(\theta \mid I)}{d\theta}$$

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

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

uniform distribution: the least informative prior

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

final result is a beta distribution

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

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

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

beta function

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

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

Mathematical digression: relationship between gamma and beta function

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

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

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

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

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

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

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

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

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

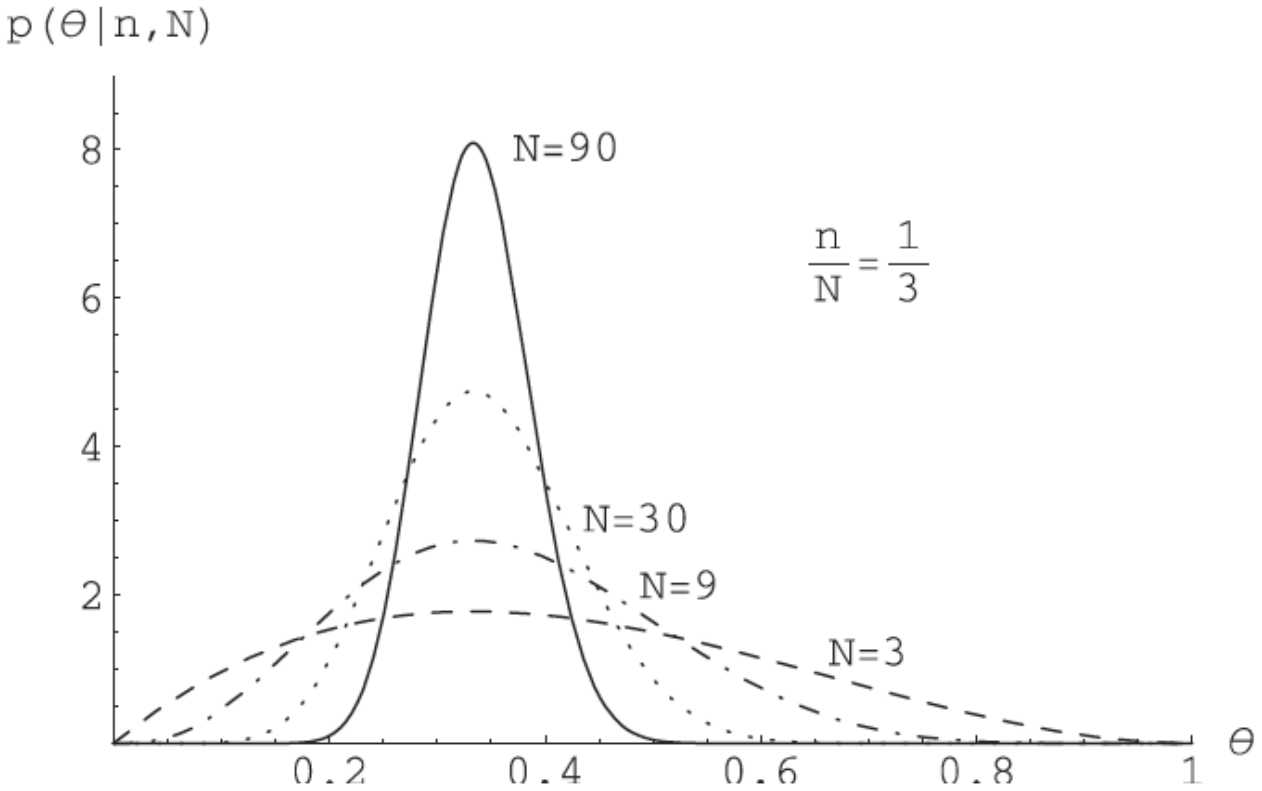


Figure 1. Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

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

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

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

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

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

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

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

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

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

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

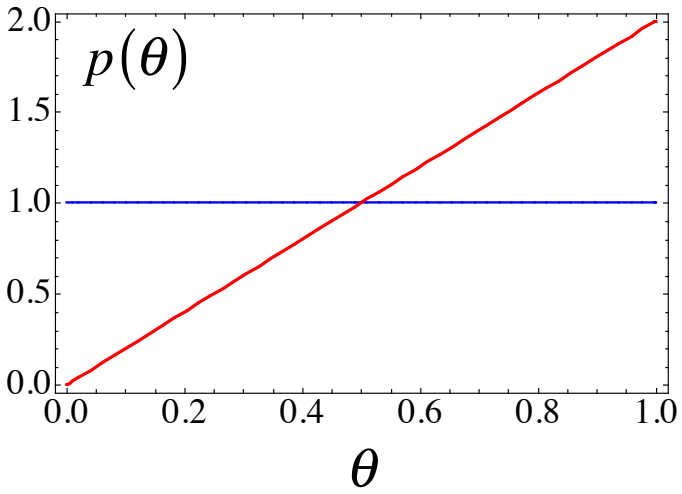
$$\operatorname{var}\theta = \left\langle \theta^2 \right\rangle - \left\langle \theta \right\rangle^2 = \frac{(n+2)(n+1)}{(N+3)(N+2)} - \left(\frac{n+1}{N+2}\right)^2 =$$

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

What happens if we try a different prior?

Let's try with a linear prior

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



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

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

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

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

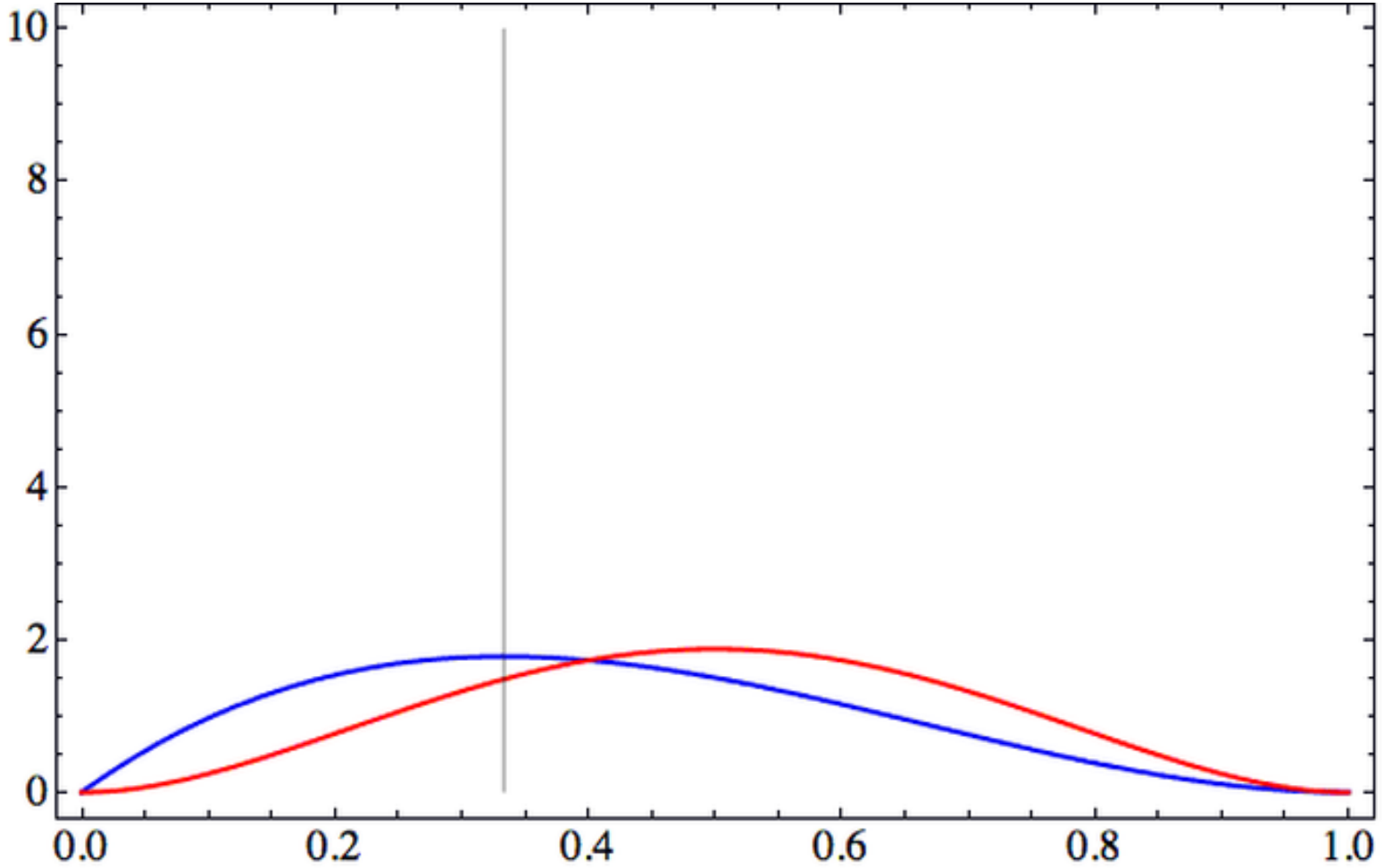
$$= \frac{(N+2)!}{(n+1)!(N-n)!}B(n+3,N-n+1)$$

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

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

Blue: uniform prior

Red: linear prior

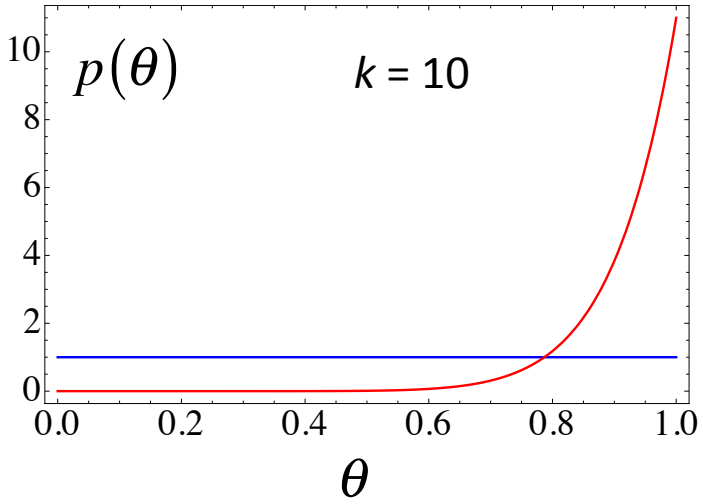


Taking few coin throws, the posterior from the linear prior is considerably biased. The bias disappears when the number of coin throws is large.

Now we try with a very non-uniform prior

We take

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



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

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

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

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

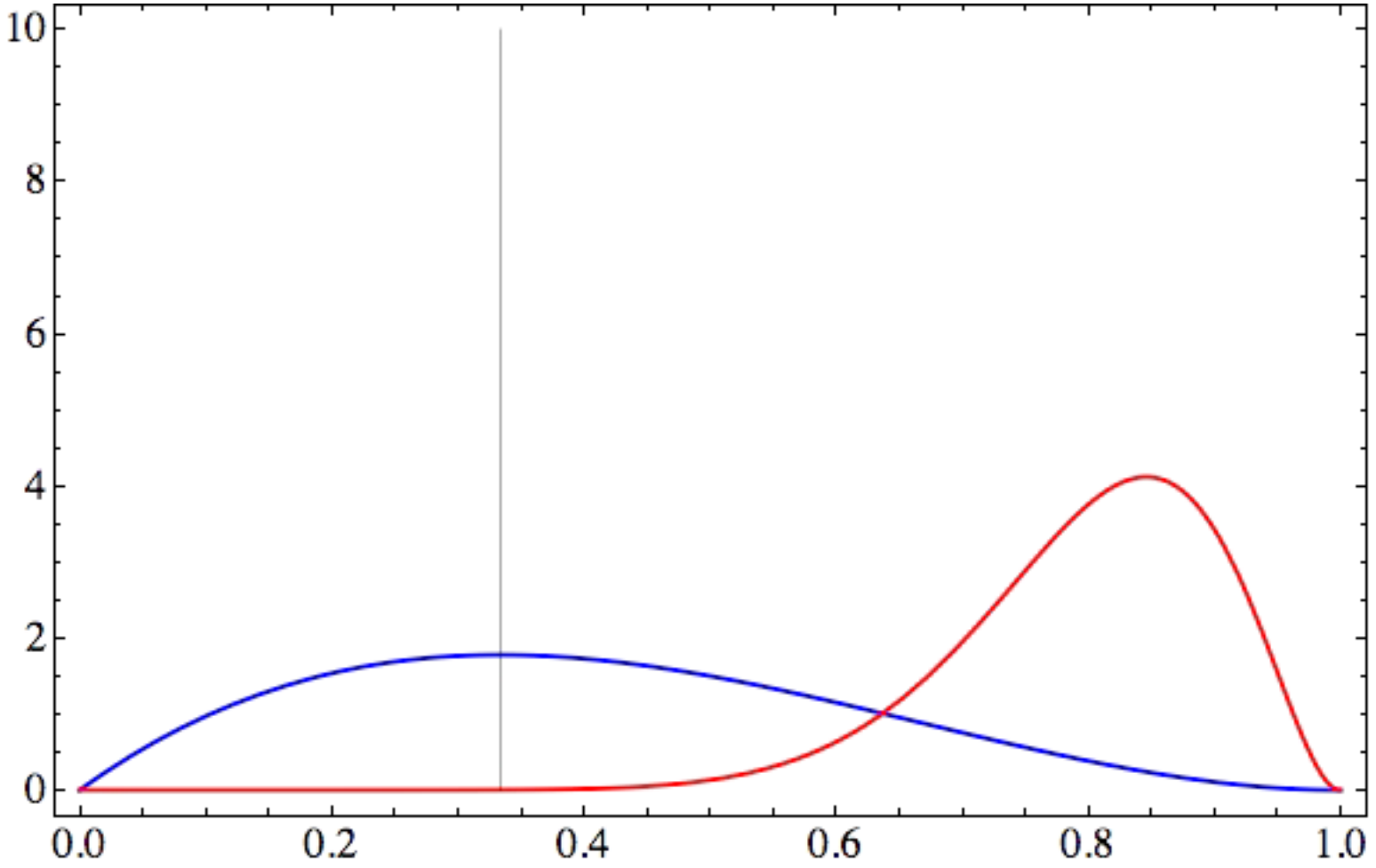
$$= \frac{(N+k+1)!}{(n+k)!(N-n)!}B(n+k+2,N-n+1)$$

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

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

Blue: uniform prior

Red: power-law prior (k=10)



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

Final note:

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

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

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

2. *Example of Bayesian inference*: estimate of a Poissonian rate

$$P(n \mid \lambda, \Delta t) = \frac{(\lambda \Delta t)^n e^{-\lambda \Delta t}}{n!}$$

uniform distribution (step function, not normalizable, improper prior!)

$$p(\lambda \mid n, \Delta t) = \frac{P(n \mid \lambda, \Delta t)}{\int_\lambda P(n \mid \lambda, \Delta t) \cdot p(\lambda) d\lambda} \cdot p(\lambda) =$$

$$= \frac{(\lambda \Delta t)^n e^{-\lambda \Delta t}}{\int_0^\infty (\lambda \Delta t)^n e^{-\lambda \Delta t} d(\lambda \Delta t)} = \frac{(\lambda \Delta t)^n e^{-\lambda \Delta t}}{n!}$$

here the random variable is the parameter λ

$$\langle \lambda\Delta t \rangle = \int_0^\infty p(\lambda \mid n, \Delta t)(\lambda\Delta t)\, d(\lambda\Delta t) = \int_0^\infty \frac{(\lambda\Delta t)^n e^{-\lambda\Delta t}}{n!}(\lambda\Delta t)\, d(\lambda\Delta t)$$

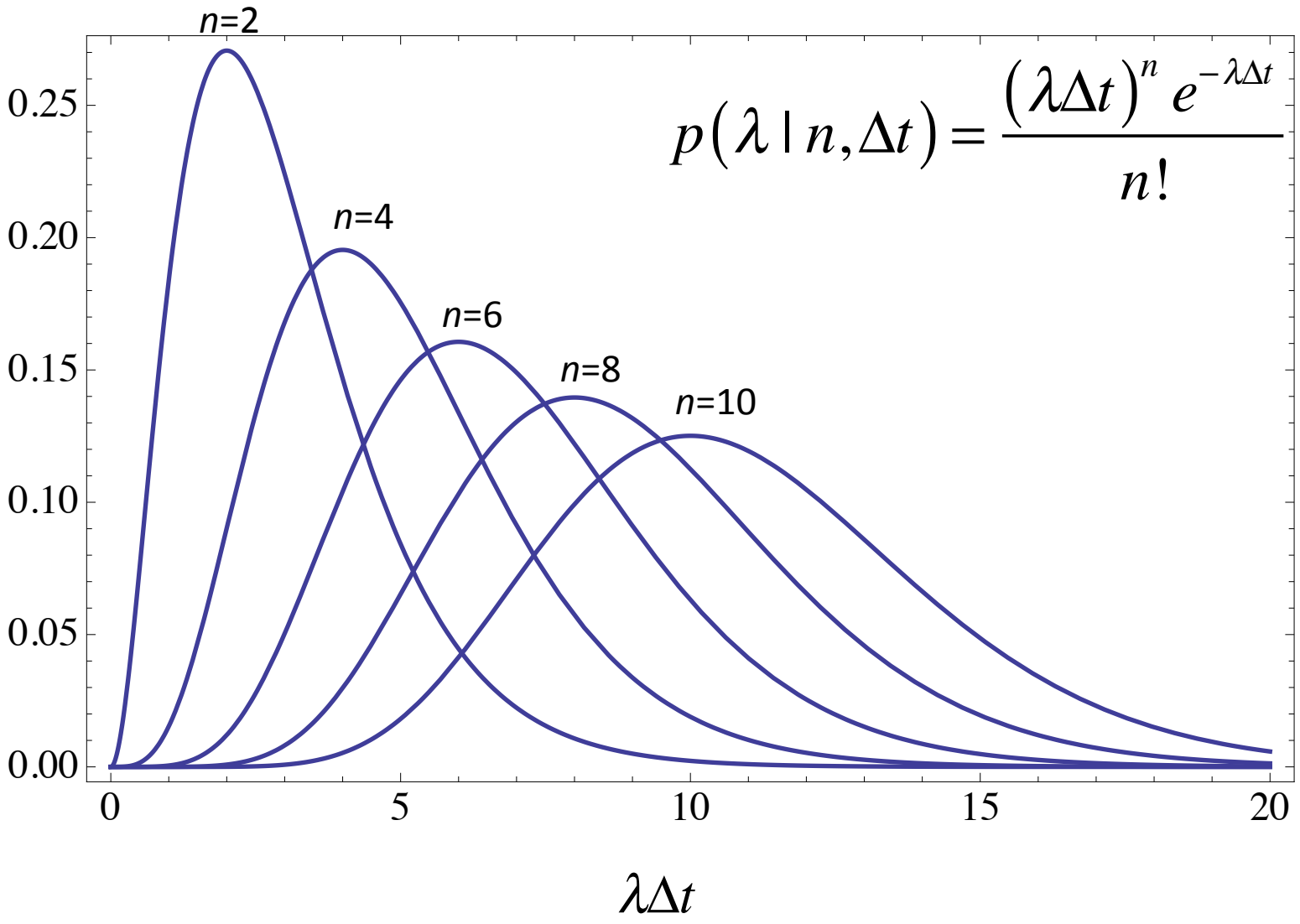
$$= \frac{1}{n!}\int_0^\infty (\lambda\Delta t)^{n+1} e^{-\lambda\Delta t} d(\lambda\Delta t) = \frac{(n+1)!}{n!} = n+1$$

$$\left\langle (\lambda\Delta t)^2 \right\rangle = \int_0^\infty \frac{(\lambda\Delta t)^n e^{-\lambda\Delta t}}{n!}(\lambda\Delta t)^2\, d(\lambda\Delta t)$$

$$= \frac{1}{n!}\int_0^\infty (\lambda\Delta t)^{n+2} e^{-\lambda\Delta t} d(\lambda\Delta t) = \frac{(n+2)!}{n!} = (n+2)(n+1)$$

$$\mathrm{var}(\lambda\Delta t) = \left\langle (\lambda\Delta t)^2 \right\rangle - \langle \lambda\Delta t \rangle^2 = (n+2)(n+1) - (n+1)^2 = n+1$$

Posterior distributions of the adimensional parameter $\lambda\Delta t$ when n events are observed



3. *Example of Bayesian inference*: Gaussian model (estimate of the mean with given distribution variance)

$$p(d \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(d-\mu)^2}{2\sigma^2} \right]$$

1. we know nothing about the position of the mean

uniform prior (improper!)

$$p(\mu \mid d,\sigma) = \frac{P(d \mid \mu,\sigma)}{\int_\mu P(d \mid \mu,\sigma) \cdot p(\mu) d\mu} \cdot p(\mu) =$$

$$= \frac{\exp\left[-(d-\mu)^2 / 2\sigma^2 \right]}{\int_{-\infty}^{+\infty} \exp\left[-(d-\mu)^2 / 2\sigma^2 \right] d\mu} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(d-\mu)^2}{2\sigma^2} \right]$$

here the random variable is μ

2. positive mean (now we know that μ > 0)

step function (another improper prior)

$$p(\mu \mid d,\sigma) = \frac{P(d \mid \mu,\sigma)}{\int_{\mu} P(d \mid \mu,\sigma)\cdot p(\mu)\,d\mu}\cdot p(\mu) =$$

$$= \frac{\exp\left[-(d-\mu)^2/2\sigma^2\right]}{\int_0^{+\infty} \exp\left[-(d-\mu)^2/2\sigma^2\right]d\mu}\cdot s(\mu) = N\exp\left[-\frac{(d-\mu)^2}{2\sigma^2}\right]\cdot s(\mu)$$

normalization

NB: because of the step function, posteriors always vanish for μ < 0

3. Gaussian prior

$$P(\mu \mid \mu_1, I) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(\mu - \mu_1)^2}{2\sigma_1^2} \right]$$

this is an informative prior, we know that the mean is centered about μ_1

$$p(\mu \mid d, \sigma) = \frac{P(d \mid \mu, \sigma)}{\int_\mu P(d \mid \mu, \sigma) \cdot p(\mu) d\mu} \cdot p(\mu) =$$

$$= \frac{\exp\left[-\frac{(d-\mu)^2}{2\sigma^2} \right]}{\int_{-\infty}^{+\infty} \exp\left[-\frac{(d-\mu)^2}{2\sigma^2} \right] \exp\left[-\frac{(\mu-\mu_1)^2}{2\sigma_1^2} \right] d\mu} \cdot \exp\left[-\frac{(\mu-\mu_1)^2}{2\sigma_1^2} \right]$$

$$p(\mu \mid d,\sigma) = \frac{\exp\left[-\frac{(d-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_1)^2}{2\sigma_1^2}\right]}{\int_{-\infty}^{+\infty} \exp\left[-\frac{(d-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_1)^2}{2\sigma_1^2}\right] d\mu}$$

this is a Gaussian expression and we can rearrange the exponent

$$\frac{(d-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_1)^2}{\sigma_1^2} = \frac{\sigma_1^2(d-\mu)^2 + \sigma^2(\mu-\mu_1)^2}{\sigma^2\sigma_1^2} = \frac{\left(\sigma_1^2+\sigma^2\right)\mu^2 - 2\left(\sigma_1^2 d + \sigma^2\mu_1\right)\mu + \left(\sigma_1^2 d^2 + \sigma^2\mu_1^2\right)}{\sigma^2\sigma_1^2}$$

$$= \frac{1}{\sigma^2\sigma_1^2}\left[\left(\sigma_1^2+\sigma^2\right)\mu^2 - 2\frac{\left(\sigma_1^2 d + \sigma^2\mu_1\right)}{\sqrt{\sigma_1^2+\sigma^2}}\sqrt{\sigma_1^2+\sigma^2}\,\mu + \frac{\left(\sigma_1^2 d + \sigma^2\mu_1\right)^2}{\left(\sigma_1^2+\sigma^2\right)} - \frac{\left(\sigma_1^2 d + \sigma^2\mu_1\right)^2}{\left(\sigma_1^2+\sigma^2\right)} + \left(\sigma_1^2 d^2 + \sigma^2\mu_1^2\right)\right]$$

$$= \frac{1}{\sigma^2\sigma_1^2}\left[\left(\sigma_1^2+\sigma^2\right)\left(\mu - \frac{\left(\sigma_1^2 d + \sigma^2\mu_1\right)}{\left(\sigma_1^2+\sigma^2\right)}\right)^2 - \frac{\left(\sigma_1^2 d + \sigma^2\mu_1\right)^2}{\left(\sigma_1^2+\sigma^2\right)} + \left(\sigma_1^2 d^2 + \sigma^2\mu_1^2\right)\right]$$

$$p(\mu \mid d,\sigma) = \frac{\exp\left\{-\frac{1}{2\sigma^2\sigma_1^2}\left[\left(\sigma_1^2+\sigma^2\right)\left(\mu-\frac{\left(\sigma_1^2 d+\sigma^2\mu_1\right)}{\left(\sigma_1^2+\sigma^2\right)}\right)^2-\frac{\left(\sigma_1^2 d+\sigma^2\mu_1\right)^2}{\left(\sigma_1^2+\sigma^2\right)}+\left(\sigma_1^2 d^2+\sigma^2\mu_1^2\right)\right]\right\}}{\int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2\sigma^2\sigma_1^2}\left[\left(\sigma_1^2+\sigma^2\right)\left(\mu-\frac{\left(\sigma_1^2 d+\sigma^2\mu_1\right)}{\left(\sigma_1^2+\sigma^2\right)}\right)^2-\frac{\left(\sigma_1^2 d+\sigma^2\mu_1\right)^2}{\left(\sigma_1^2+\sigma^2\right)}+\left(\sigma_1^2 d^2+\sigma^2\mu_1^2\right)\right]\right\}d\mu}$$

$$= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\sigma^2\sigma_1^2}{\left(\sigma_1^2+\sigma^2\right)}}\exp\left\{-\frac{\left(\sigma_1^2+\sigma^2\right)}{2\sigma^2\sigma_1^2}\left(\mu-\frac{\left(\sigma_1^2 d+\sigma^2\mu_1\right)}{\left(\sigma_1^2+\sigma^2\right)}\right)^2\right\}$$

$$\mu_2 = \frac{\sigma_1^2 d+\sigma^2\mu_1}{\sigma_1^2+\sigma^2}$$

$$\sigma_2^2 = \frac{\sigma^2\sigma_1^2}{\sigma_1^2+\sigma^2} \qquad \left(\frac{1}{\sigma_2^2}=\frac{1}{\sigma^2}+\frac{1}{\sigma_1^2}\right)$$

Recursivity of inferential process

$$\sigma_2^2 = \frac{\sigma^2 \sigma_1^2}{\sigma_1^2 + \sigma^2} \qquad \left(\frac{1}{\sigma_2^2} = \frac{1}{\sigma^2} + \frac{1}{\sigma_1^2} \right)$$

$$\mu_2 = \frac{d/\sigma^2 + \mu_1/\sigma_1^2}{1/\sigma^2 + 1/\sigma_1^2}$$

weighted mean

$$\mu_2 = \frac{\sigma_1^2 d + \sigma^2 \mu_1}{\sigma_1^2 + \sigma^2} = \frac{\sigma_1^2 d - \sigma_1^2 \mu_1 + \left(\sigma^2 + \sigma_1^2\right)\mu_1}{\sigma_1^2 + \sigma^2} = \mu_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2}\left(d - \mu_1\right)$$

same structure as a Kalman filter

Kalman filter (super-simple example)

$$\mu_2 = \mu_1 + K\left(d_2 - \mu_1\right)$$

$$K = \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2}$$ (*guadagno* del filtro)

+ linear variation (with Gaussian noise)

$$\mu\left(t_3\right) = \mu\left(t_2\right) + v\left(t_3 - t_2\right)$$

$$\sigma_\mu^2\left(t_3\right) = \sigma_\mu^2\left(t_2\right) + \sigma_v^2\left(t_3 - t_2\right)$$

$$\mu_3 = \mu_2 + K_3\left(d_3 - \mu_2\right)$$

$$K_3 = \frac{\sigma_\mu^2\left(t_3\right)}{\sigma_\mu^2\left(t_3\right) + \sigma_3^2}$$

Bayesian inference and maximum-likelihood

uniform distribution (in general, improper)

$$p(\boldsymbol{\theta} \mid \mathbf{d}, I) = \frac{P(\mathbf{d} \mid \boldsymbol{\theta}, I)}{P(\mathbf{d} \mid I)} \cdot p(\boldsymbol{\theta} \mid I)$$

likelihood

$$= \frac{\mathcal{L}(\mathbf{d}, \boldsymbol{\theta})}{P(\mathbf{d} \mid I)} \cdot p(\boldsymbol{\theta} \mid I) \propto \mathcal{L}(\mathbf{d}, \boldsymbol{\theta})$$

evidence

in this case the set of parameters that maximizes the posterior (MAP) is also the set that maximizes the likelihood (MLE)

References:

- G. D'Agostini, Rep. Prog. Phys. **66** (2003) 1383
- V. Dose, Rep. Prog. Phys. **66** (2003) 1421
- J. Skilling, J. of Microscopy **190** (1998) 28
- P. S. Maybeck, *Stochastic models, estimation, and control (vol. 1)*, Academic Press 1979