Introduction to Bayesian Methods- 2

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Likelihood

Prior distribution

Posterior distribution

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

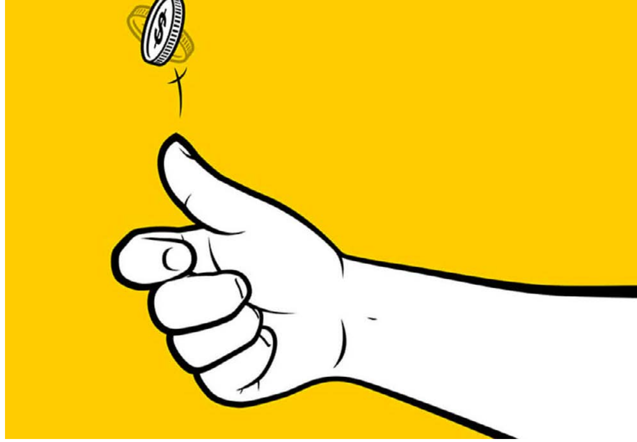
Evidence

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

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

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)

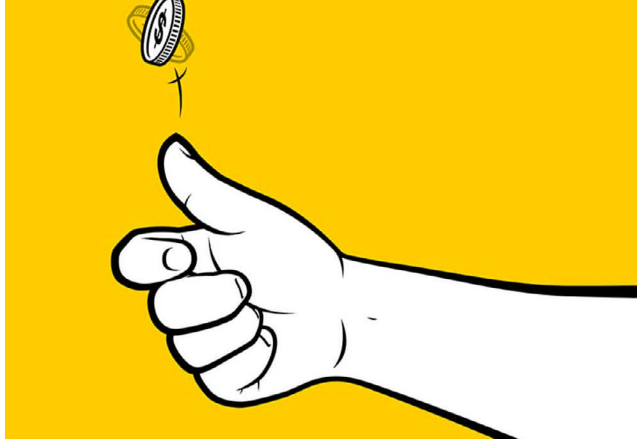
Is this an unbiased coin?

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

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

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

uniform distribution: the least informative prior

the final result is a beta distribution

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

$$\begin{aligned} B(m,n) &= \int_0^1 t^{m-1}(1-t)^{n-1}dt \\ &= \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \end{aligned}$$

beta function

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

Mathematical digression: the connection between gamma and beta function

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

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

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

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

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

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

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

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

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

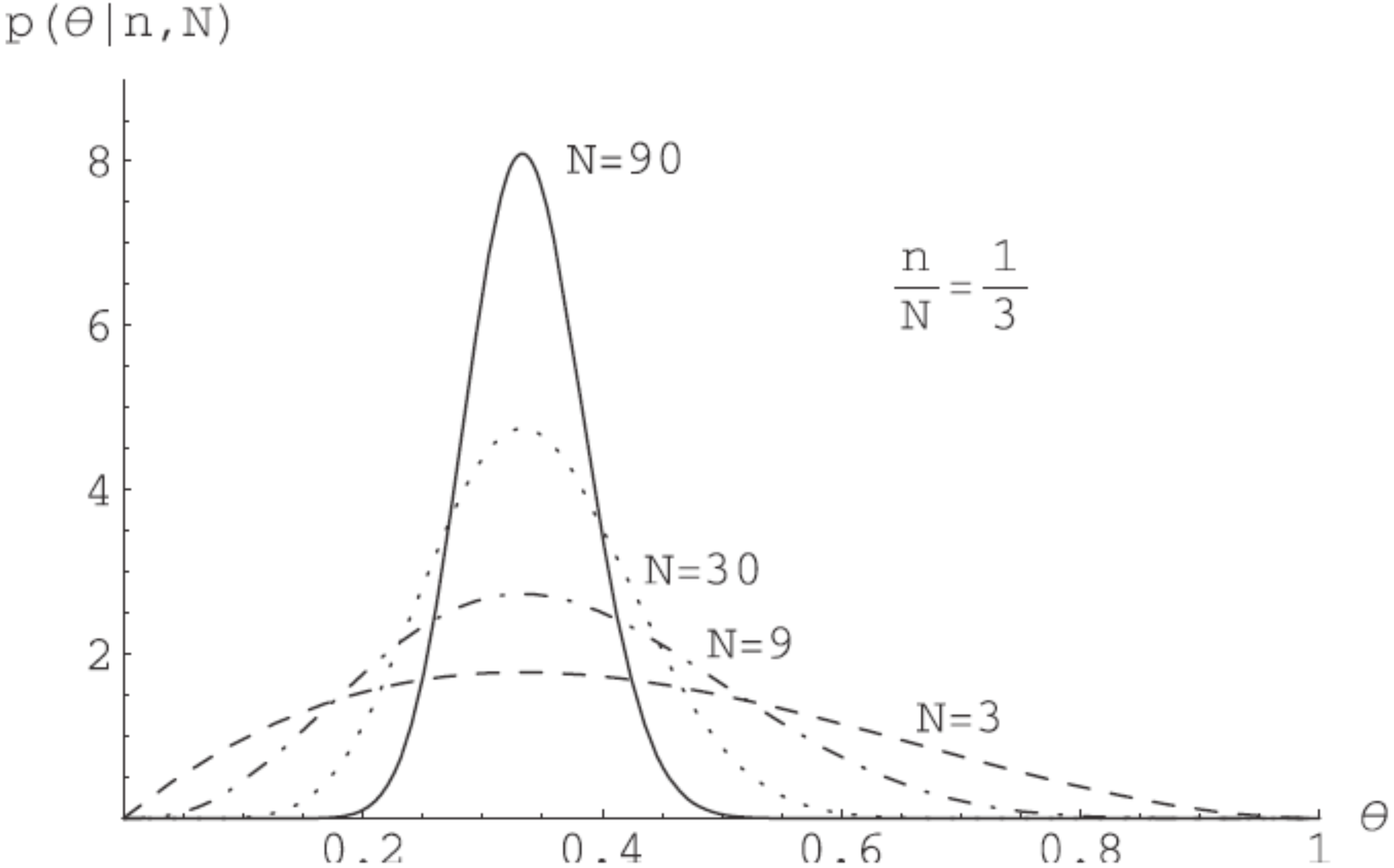


Figure 1. Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

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

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

$$\Downarrow$$

$$\begin{aligned}
\langle\theta\rangle &= \int_0^1 p\left(\theta \mid n,N\right)\theta d\theta = \frac{(N+1)!}{n!(N-n)!}\int_0^1 \left(1-\theta\right)^{N-n}\theta^{n+1}d\theta \\
&= \frac{(N+1)!}{n!(N-n)!}B\left(n+2,N-n+1\right) \\
&= \frac{(N+1)!}{n!(N-n)!}\cdot\frac{(n+1)!(N-n)!}{(N+2)!} \\
&= \frac{n+1}{N+2} \to \frac{n}{N}
\end{aligned}$$

biased, asymptotically unbiased, estimator

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

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

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

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

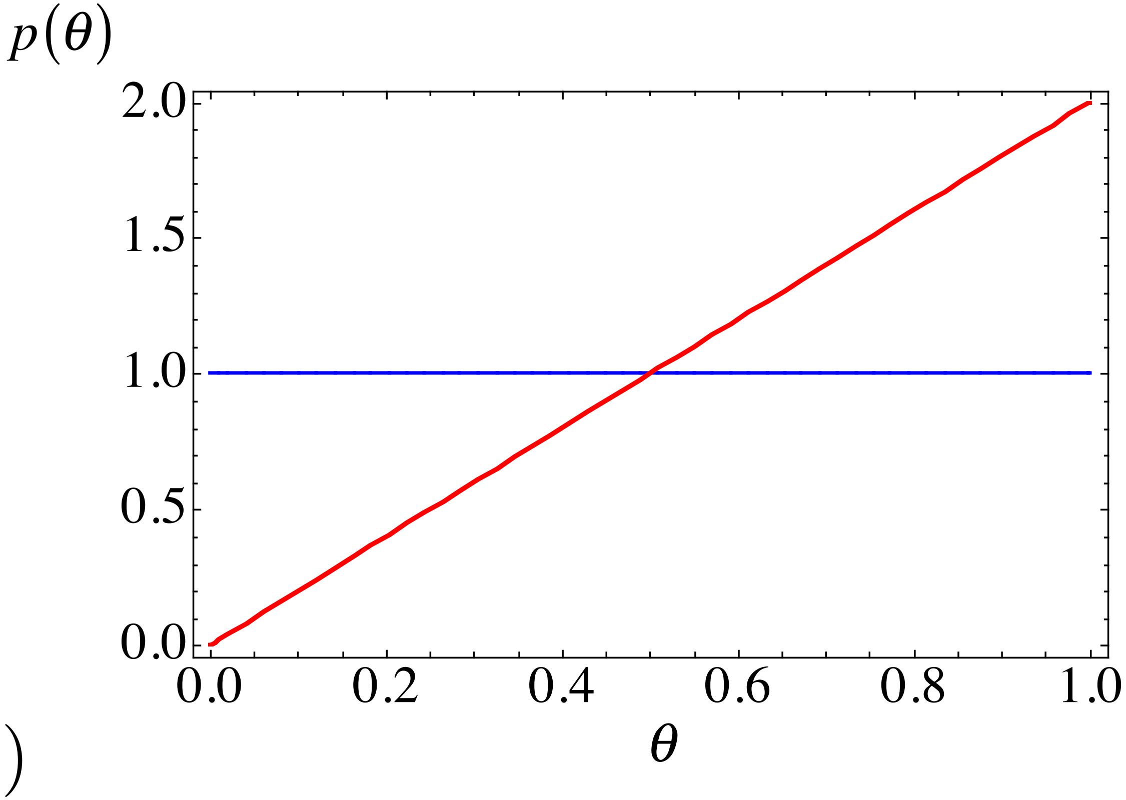
$$\operatorname{var}\theta = \left\langle \theta^2 \right\rangle - \left\langle \theta \right\rangle^2 = \frac{\left(n+2\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)} - \left(\frac{n+1}{N+2}\right)^2 =$$

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

What happens if we try a different prior?

Let's try with a linear prior

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



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

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

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

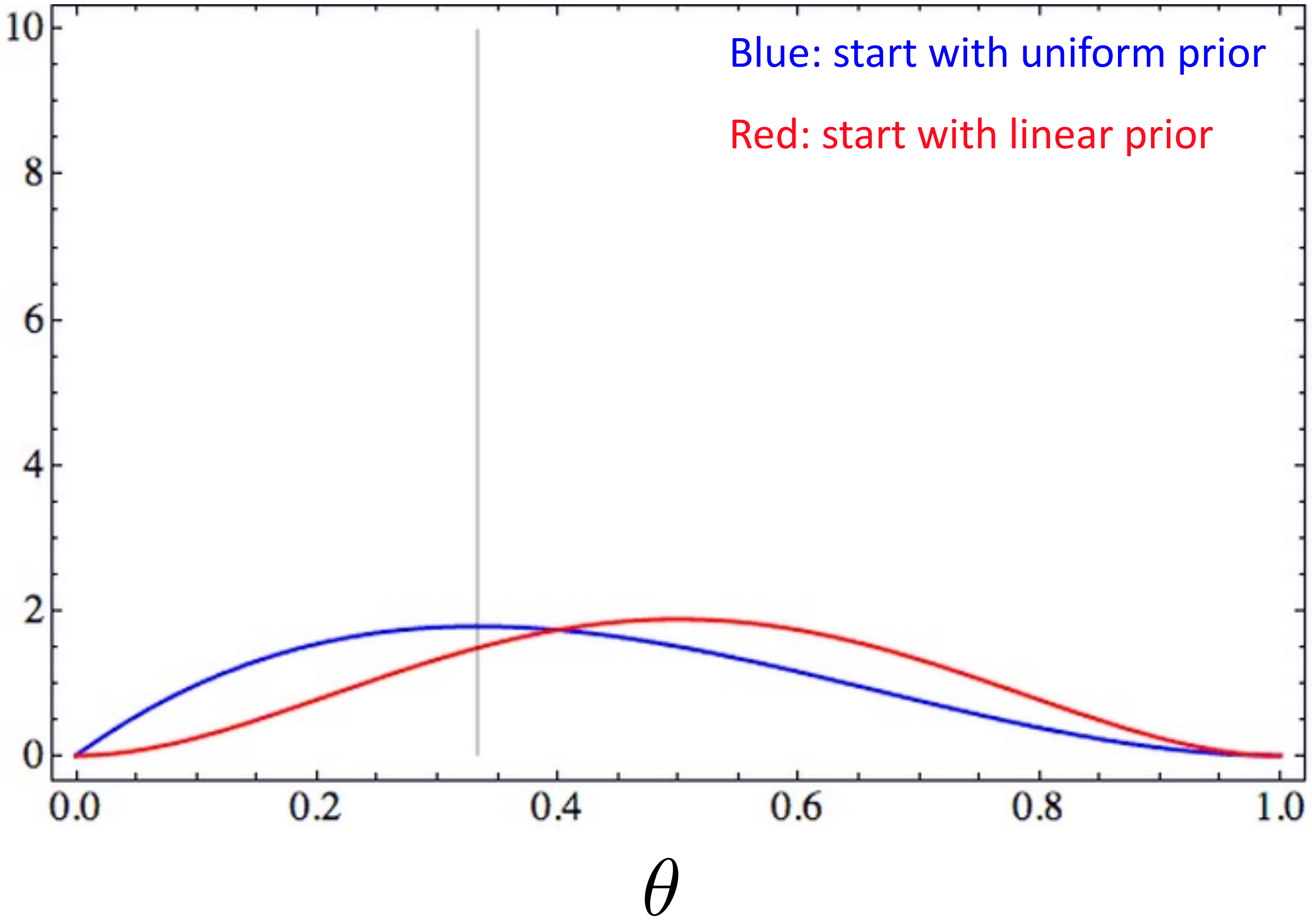
$$\Downarrow$$

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

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

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

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

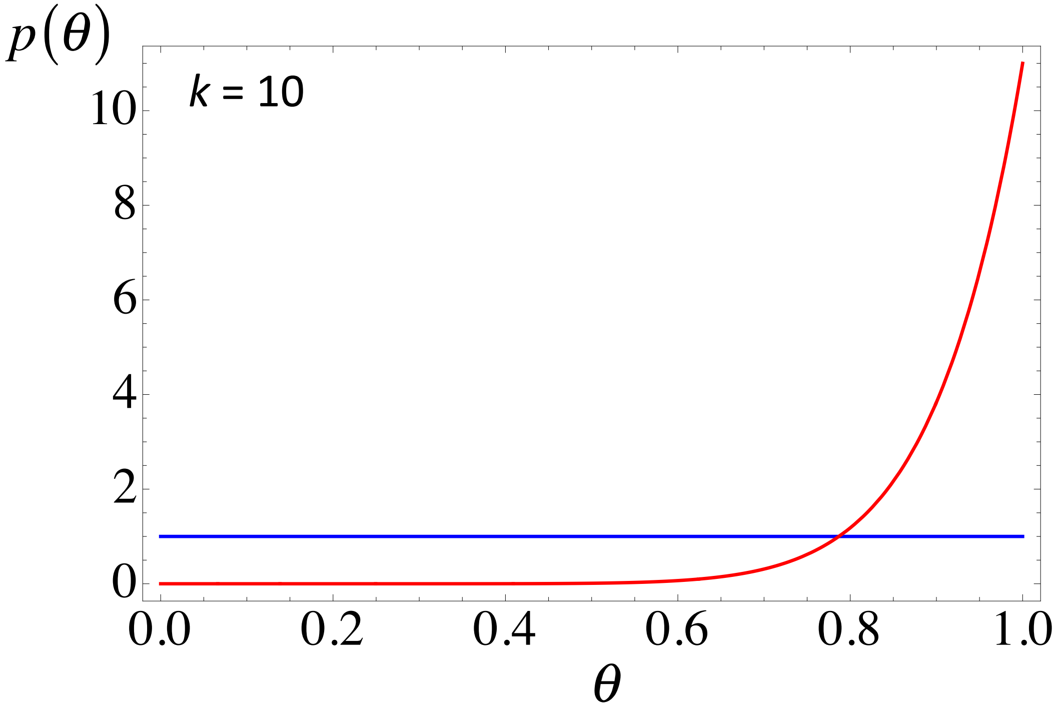


Taking few coin throws, the posterior from the linear prior is considerably biased. The bias disappears when the number of coin throws is large.

Now we try with a very non-uniform prior

We take

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



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

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

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

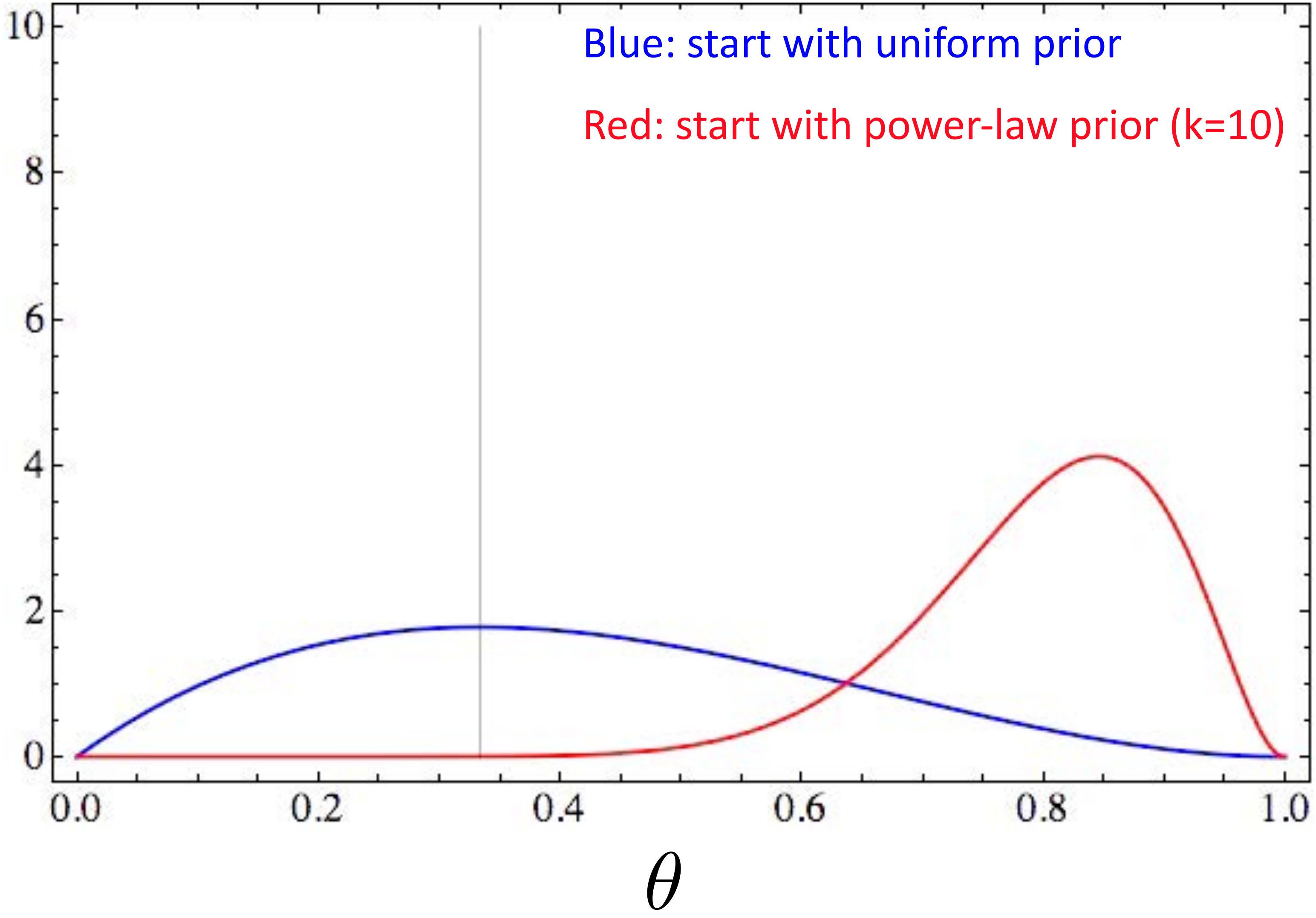
$$\Downarrow$$

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

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

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

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



Blue: start with uniform prior

Red: start with power-law prior (k=10)

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 corresponds to the formal definition of the Beta distribution:

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

There are other important dualities between distributions.

We shall soon meet additional dualities for important distributions.

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

- 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.

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

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

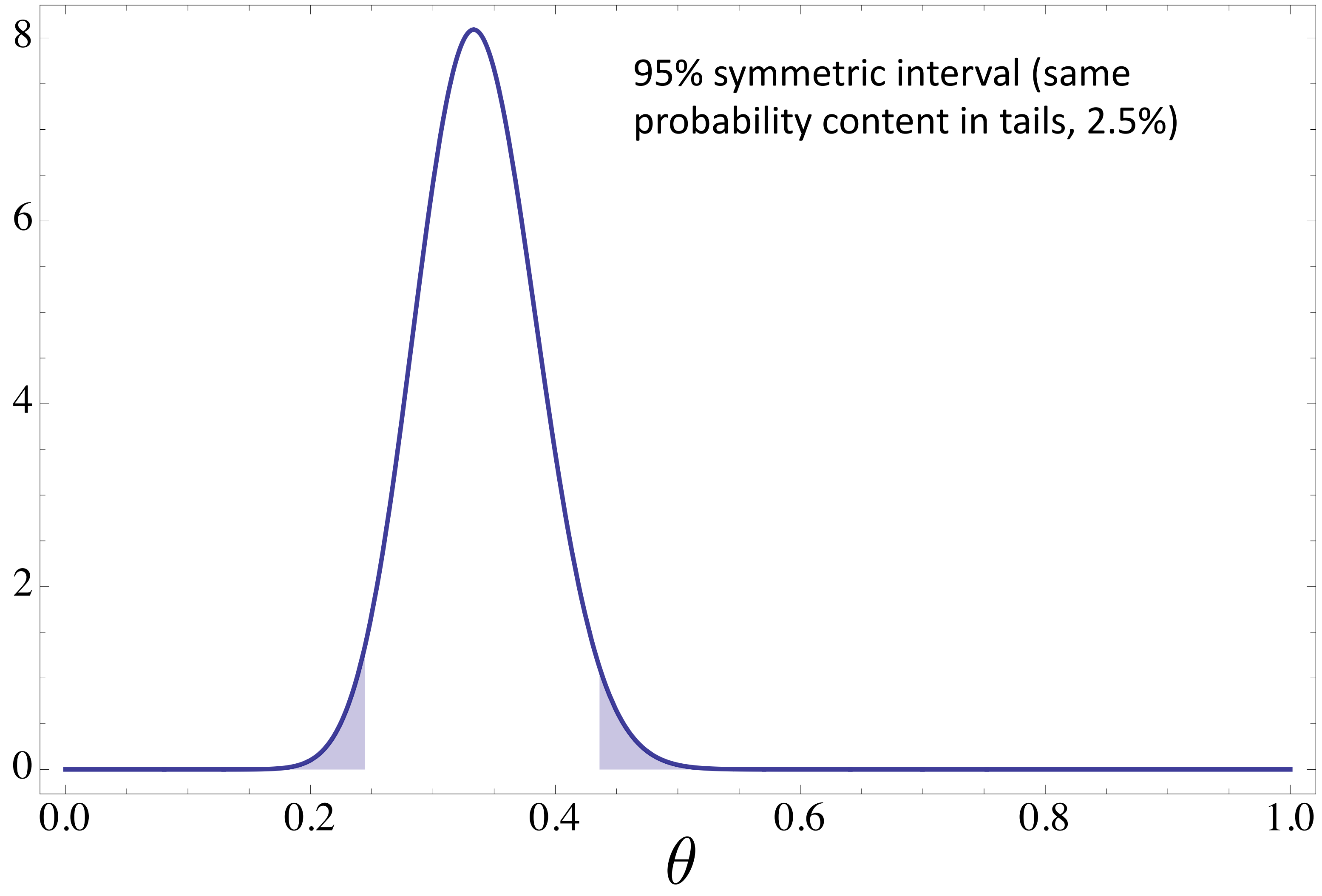
$$p(\theta \mid n, N) = \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\,\theta^n$$

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.

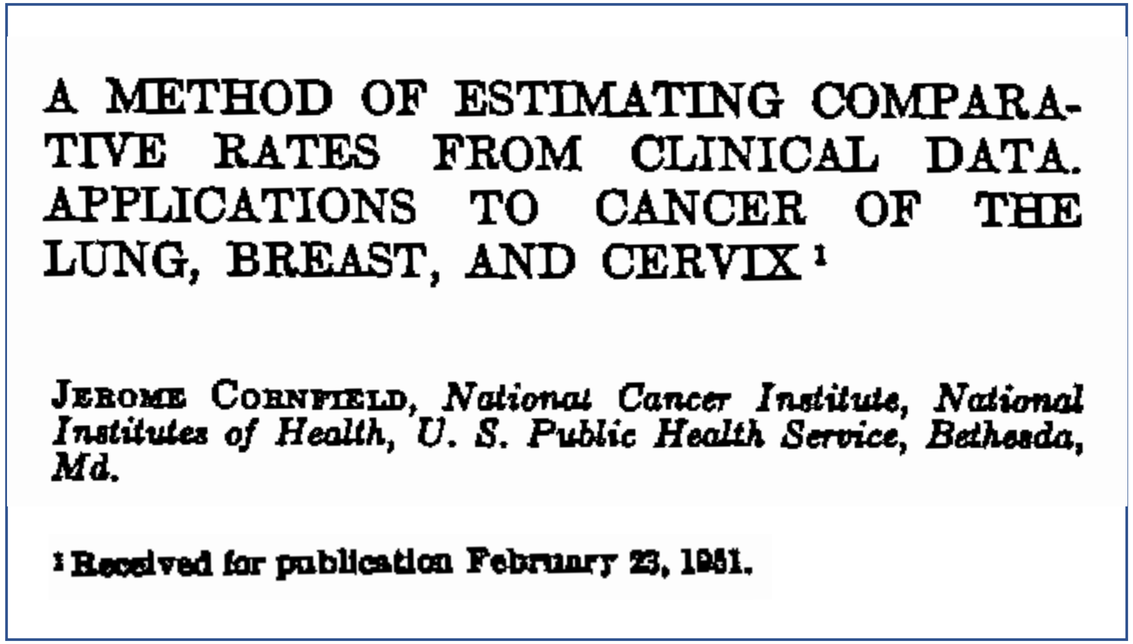


95% symmetric interval (same probability content in tails, 2.5%)

Example: the statistical link between smoking and lung cancer

Cornfield, Jerome

Born: October 30, 1912, in New York City, New York.
Died: September 17, 1979, in Herndon, Virginia.



A METHOD OF ESTIMATING COMPARATIVE RATES FROM CLINICAL DATA. APPLICATIONS TO CANCER OF THE LUNG, BREAST, AND CERVIX [1]

JEROME CORNFIELD, *National Cancer Institute, National Institutes of Health, U. S. Public Health Service, Bethesda, Md.*

[1] Received for publication February 23, 1951.

Cornfield, Jerome

Born: October 30, 1912, in New York City, New York.
Died: September 17, 1979, in Herndon, Virginia.

Jerome Cornfield was arguably the most influential statistician in the biomedical sciences in the US from the 1950s until his death. He was the consummate statistical scientist. His understanding of the nature of the subject-matter of statistics and of its essential role in the inductive process of integrating data into a body of empirical knowledge, particularly in the biomedical sciences, was outstanding. This thorough view of statistics and scientific research enabled him to identify essential statistical problems. He exercised considerable influence as an advisor and consultant, and for over two decades was a major advocate for statistical reasoning in clinical research.

After attending elementary and high schools in the Bronx, New York, he entered New York University, graduating in 1933 with a major in history. Cornfield did not receive any advanced degrees. He did, however, take some formal graduate courses in history at Columbia University. After moving to Washington, DC, in 1935, Cornfield took a number of courses in statistics at the US Department of Agriculture Graduate School during the period 1936 – 1938, including courses with M.A. Girshick in general statistics and **multivariate analysis**.

...

Cornfield, Jerome

Born: October 30, 1912, in New York City, New York.
Died: September 17, 1979, in Herndon, Virginia.

...

Over a span of three decades, from 1947 to 1979, Professor Cornfield was one of the leading statisticians working in the biomedical area. He made many original contributions to biostatistics, epidemi- ology, **clinical trials**, and to quantitative methods in the design and analysis of experiments conducted in clinical and laboratory research. In addition, he wrote a number of papers on Bayesian **inference** and on the application of **Bayesian methods** in the biomedical sciences.

...

From 1948 to his death 31 years later, Cornfield devoted the major portion of his career to the development and application of statistical theory and methods to the biomedical sciences. His contributions were diverse both in the nature of his statistical interests and in the areas of biostatistical applications. He was involved in and touched upon every major public health issue that arose in that period – the polio vaccines, smoking and lung cancer, risk factors for cardiovascular disease, and the difficult statistical issues of estimating the low-dose carcinogenic effects in humans of a food additive that becomes suspect because it produces cancer in animals at much higher doses.

...

(excerpted from *Encyclopedia of Biostatistics*, Online © 2005 John Wiley & Sons, Ltd.
This article DOI: 10.1002/0470011815.b2a17032)

FIGURE 1. Passport photograph of Ronald Aylmer Fisher at age 34. Reprinted from Box JF. RA Fisher: the life of a scientist. New York: John Wiley & Sons, Inc., 1978.

Fisher developed four lines of argument in questioning the causal relation of lung cancer to smoking.

1) If A is associated with B, then not only is it possible that A causes B, but it is also possible that B is the cause of A. In other words, smoking may cause lung cancer, but it is a logical possibility that lung cancer causes smoking.

2) There may be a genetic predisposition to smoke (and that genetic predisposition is presumably also linked to lung cancer).

3) Smoking is unlikely to cause lung cancer because secular trend and other ecologic data do not support this relation.

4) Smoking does not cause lung cancer because inhalers are less likely to develop lung cancer than are noninhalers

Lung cancer and cigarette smoking

Consider the following data for fractions of the population (Cornfield, 1951)

	Having cancer of the lung	Healthy	Total
Smokers	$0.119 \cdot 10^{-3}$	0.579910	0.580025
Nonsmokers	$0.036 \cdot 10^{-3}$	0.419935	0.419971
Total	$0.155 \cdot 10^{-3}$	0.999845	1.000000

what is the proportion having cancer of the lung in each population?

Consider the populations of smokers (S) and non–smokers (N), and the two conditions, healthy (H) or sick with cancer (C), then using Bayes' theorem we can write:

$$P(C|S) = \frac{P(SC)}{P(S)} \qquad P(C|N) = \frac{P(NC)}{P(N)} \quad \Rightarrow \quad \frac{P(C|S)}{P(C|N)} = \frac{P(SC)/P(S)}{P(NC)/P(N)}$$

Then:

Smokers: $0.119 \cdot 10^{-3}/0.580025 = 2.05164 \cdot 10^{-4}$

Nonsmokers: $0.036 \cdot 10^{-3}/0.419971 = 8.57202 \cdot 10^{-5}$

Therefore, the prevalence of lung cancer in smokers with respect to nonsmokers is

Smokers/Nonsmokers ≈ 2.4

- In 1954 Richard Doll and Bradford Hill published evidence in the British Medical Journal showing a strong link between smoking and lung cancer. They published further evidence in 1956.

- Fisher was a paid tobacco industry consultant and a devoted pipe smoker. **He did not think the statistical evidence for a link was convincing.**

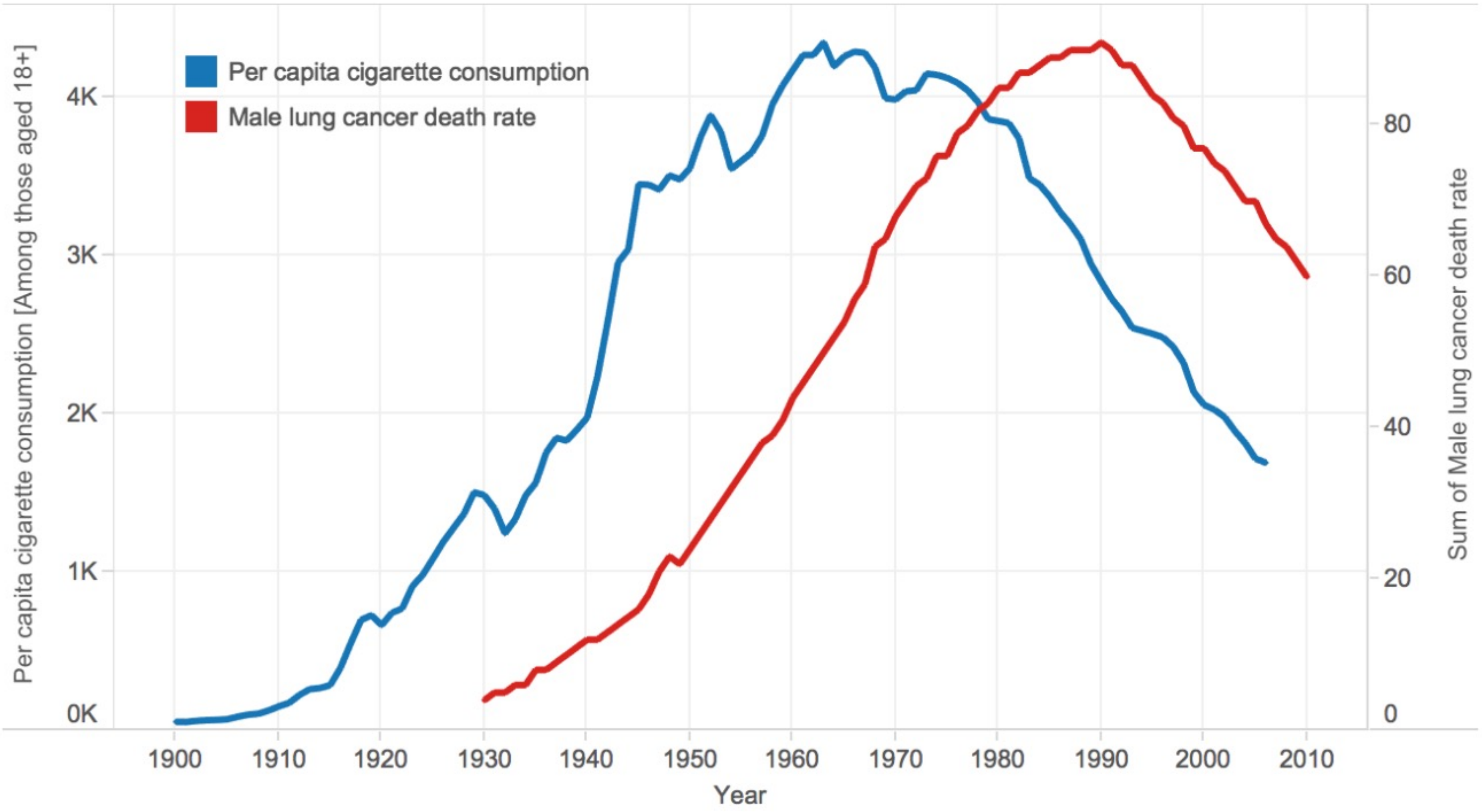
- Ronald Fisher died aged 72 on July 29, 1962, in Adelaide, Australia following an operation for colon cancer.

- With bitter irony, we now know that the likelihood of getting this disease increases in smokers.

  Ronald Fisher was cremated, and his ashes interred in St. Peter's Cathedral, Adelaide.

(from "Ronald Fisher." Famous Scientists. famousscientists.org. 17 Sep. 2015. Web. 5/30/2017 <www.famousscientists.org/ronald-fisher/>.)

Trends in Tobacco Use and Lung Cancer Death Rates in the U.S.



Death rates source: US Mortality Data, 1960-2010, US Mortality Volumes, 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention.
Cigarette consumption source: US Department of Agriculture, 1900-2007.

Example: analysis of a decision problem (Skilling 1998)

Let T be the temperature of a liquid which can be either water or ethanol. **We use the temperature data to discriminate between water and 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*

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

5. Using Bayes' theorem we find

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

* 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. $p(T) = (160°\mathrm{C})^{-1}$ in -80°C < T < 80°C.

2. $p(D \mid T, \text{ethanol}, I) = 0.1\ (°\mathrm{C})^{-1}$ in [T-5°C, T+5°C], and zero elsewhere.

3. We take a single measurement D = -3°C.

4. The evidence $p(D \mid \text{ethanol}, I)$ is

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

5. Using Bayes' theorem, we find

$$\begin{aligned} p(T|D, \text{ethanol}, I) &= \frac{p(D|T, \text{ethanol}, I)}{p(D, \text{ethanol}, I)} p(T|\text{ethanol}, I) = \frac{0.1(°\mathrm{C})^{-1}}{0.00625(°\mathrm{C})^{-1}} \frac{1}{160} (°\mathrm{C})^{-1} \\ &= 0.1(°\mathrm{C})^{-1} \qquad (-8°\mathrm{C} < T < 2°\mathrm{C}) \end{aligned}$$

- Here we only wish to discriminate between water and ethanol and we do not care much about temperature.

- Temperature is a *nuisance variable*, one that can be dispensed with.

- Usually, nuisance variable are eliminated by integration. In this specific case we have already carried out part of the work by calculating the evidences, which can be considered as marginalized likelihoods.

Assuming a uniform 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 the ratio of the posteriors is given by the Bayes' factor

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

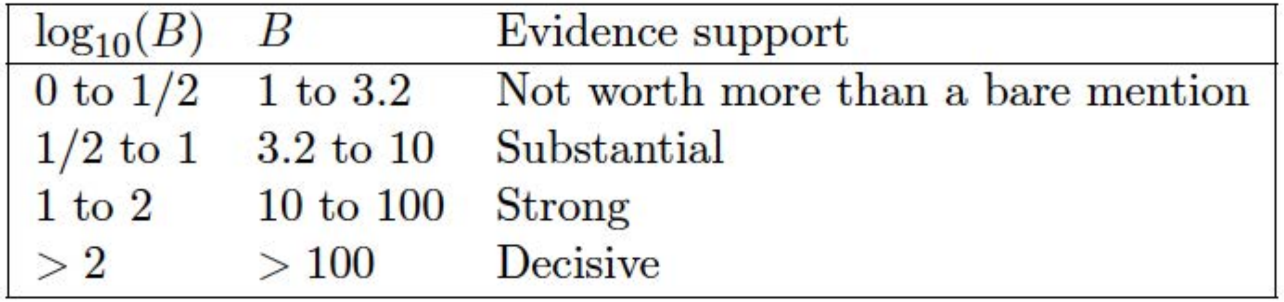
We found earlier that

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

therefore, the Bayes factor is

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

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

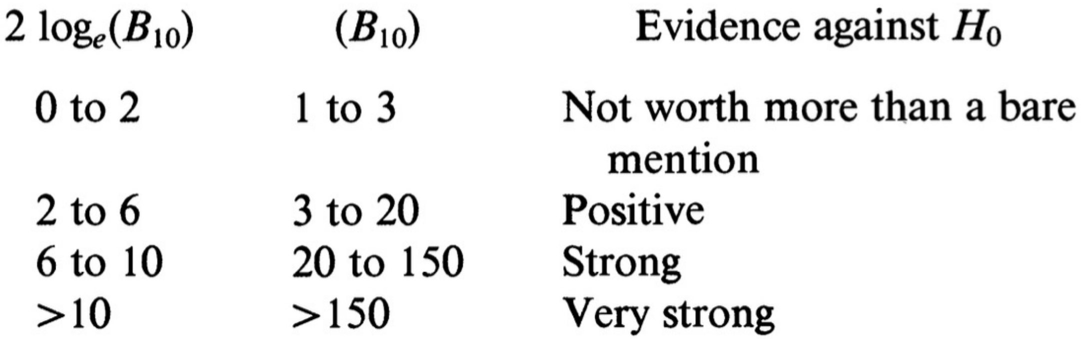


$\log_{10}(B)$	B	Evidence support
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

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 uncertainty

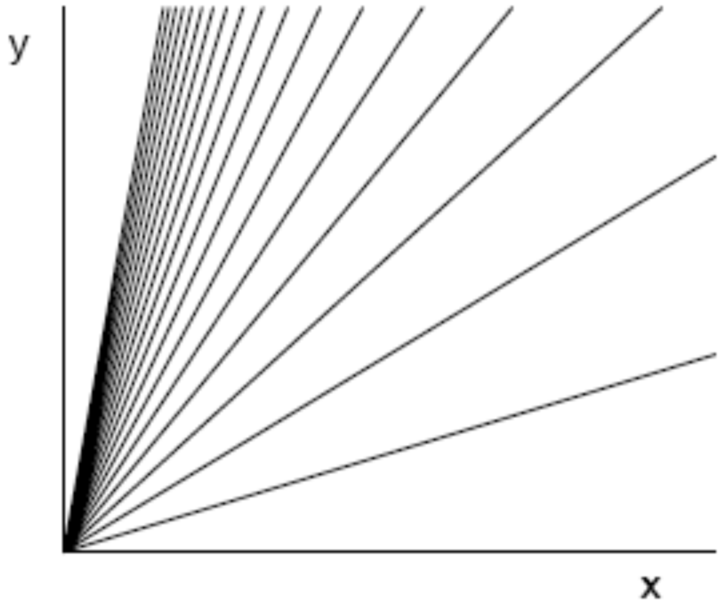
the statistical measurement uncertainty
has a Gaussian distribution

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

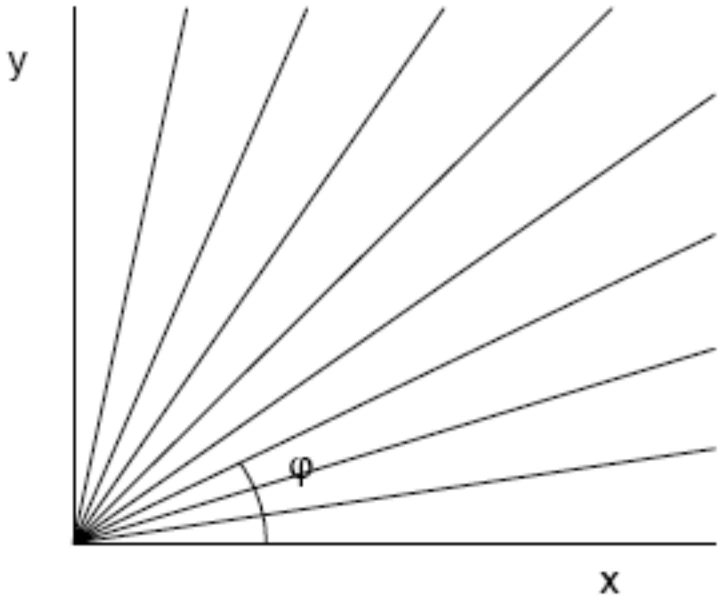
likelihood

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

prior angular distribution



uniform *a*



uniform angle

Should we take a uniform *a* or a uniform angle?

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

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

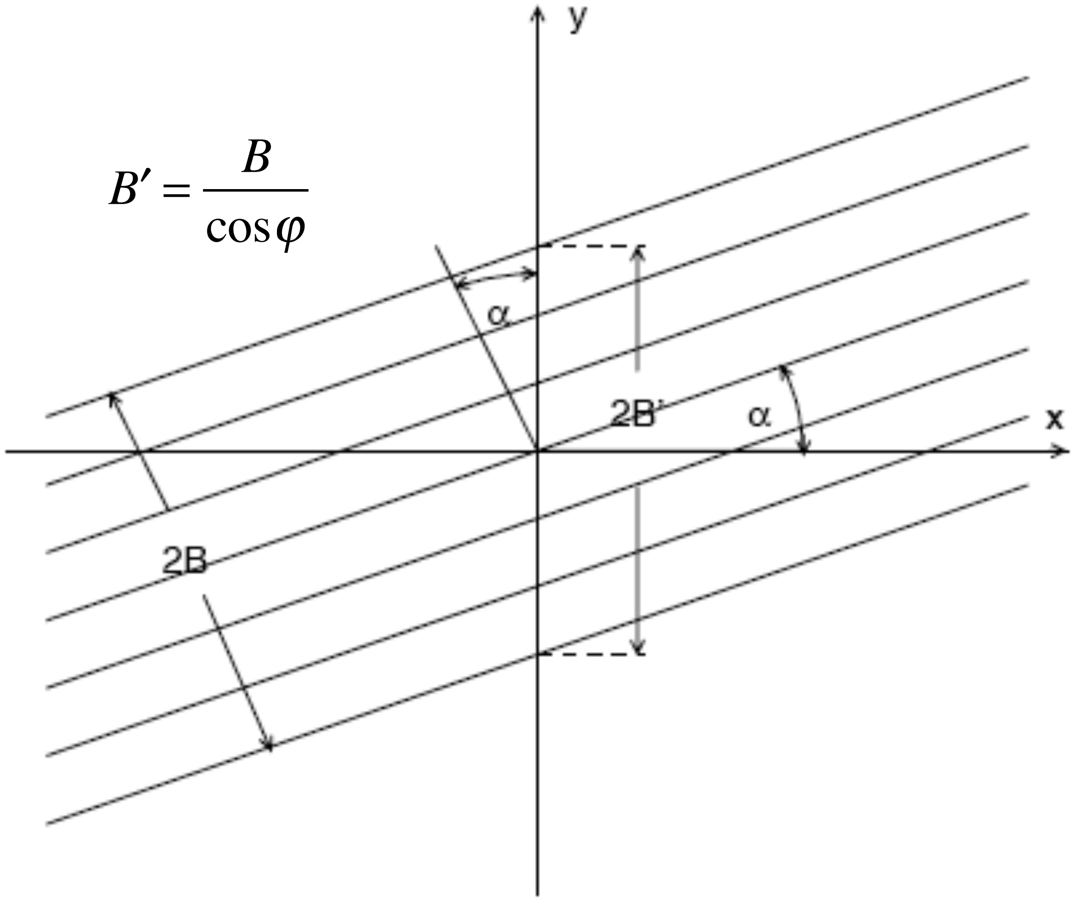
and we obtain the distribution of *a* with the transformation method:

$$a = \tan\varphi$$

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

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

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



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

finally, we obtain the posterior from Bayes' theorem

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 obtain the posterior from Bayes' theorem

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

where the prior is

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

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

therefore:

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

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

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

$$
\begin{aligned}
\sum_{i=1}^{N}\left(y_i-ax_i-b\right)^2 &= \sum_{i=1}^{N}\left[\left(y_i-ax_i\right)^2-2b\left(y_i-ax_i\right)+b^2\right] \\
&= \sum_{i=1}^{N}\left(y_i-ax_i\right)^2-2b\sum_{i=1}^{N}\left(y_i-ax_i\right)+Nb^2 \\
&= N\left\{\left[b^2-2b\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)+\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right]+\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)^2-\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right\} \\
&= N\left\{\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)^2-\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right\} \\
&= N\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2+N\left(\frac{1}{N}\sum_{i=1}^{N}y_i^2-2a\frac{1}{N}\sum_{i=1}^{N}x_iy_i+a^2\frac{1}{N}\sum_{i=1}^{N}x_i^2\right)-N\left(\frac{1}{N}\sum_{i=1}^{N}y_i-a\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 \\
&= N\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2+N\left(\operatorname{var}y-2a\operatorname{cov}(x,y)+a^2\operatorname{var}x\right)
\end{aligned}
$$

therefore, the normalization integral becomes

$$
\begin{aligned}
&\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var}y-2a\operatorname{cov}(x,y)+a^2\operatorname{var}x\right)\right]\int_{-\infty}^{+\infty}db\exp\left[-\frac{N}{2\sigma^2}\left(b-\frac{1}{N}\sum_{i=1}^{N}\left(y_i-ax_i\right)\right)^2\right] \\
&=\sqrt{\frac{2\pi\sigma^2}{N}}\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var}y-2a\operatorname{cov}(x,y)+a^2\operatorname{var}x\right)\right]
\end{aligned}
$$

For the next step we use *Laplace's method* (this is the *saddle-point method* – also called the *method of steepest descent* in the real domain) for the evaluation of the integral of a unimodal function

$$Z = \int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{+\infty} e^{\Phi(x)}dx$$

where

$$\Phi(x) = \ln p(x) \approx \ln p(x_0) - \frac{1}{2s}(x - x_0)^2$$

where x_0 is the mode and

$$\frac{1}{s} = -\frac{\partial^2 \ln p(x)}{\partial x^2}$$

therefore

$$\boxed{Z \approx \int_{-\infty}^{+\infty} p(x_0)e^{-\frac{(x-x_0)^2}{2s}}dx = p(x_0)\sqrt{2\pi s}}$$

Approximate integration of the remaining integral with Laplace's method

$$\int_{-\infty}^{+\infty} \frac{da}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)\right]$$

Taking the logarithm of the integrand, we find its maximum and we Taylor-expand about the maximum

$$\Phi(a) = -\frac{3}{2}\ln\left(1+a^2\right) - \frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)$$

$$\Phi(a) = -\frac{3}{2}\ln\left(1+a^2\right) - \frac{N}{2\sigma^2}\left(\mathrm{var}\, y - 2a\,\mathrm{cov}(x,y) + a^2\,\mathrm{var}\, x\right)$$

$$\frac{d\Phi}{da} = -\frac{3a}{1+a^2} + \frac{N}{\sigma^2}\left(\mathrm{cov}(x,y) - a\,\mathrm{var}\, x\right) = 0$$

we find *a* from this cubic equation

note that when *N*>>1 the peak is at position $a_0 \approx \frac{\mathrm{cov}(x,y)}{\mathrm{var}\, x}$

We use the Newton-Raphson method for the solution of the cubic equation:

$$f(a_0) = -\frac{3a_0}{1+a_0^2}$$

$$f'(a_0) = -3\frac{1-a_0^2}{(1+a_0^2)^2} - \frac{N}{\sigma^2}\mathrm{var}\, x \approx -\frac{N}{\sigma^2}\mathrm{var}\, x$$

then

$$\delta a_1 = -\frac{3a_0}{1+a_0^2}\frac{\sigma^2}{N \operatorname{var} x} \qquad\qquad a_1 = a_0 - \frac{3a_0}{1+a_0^2}\frac{\sigma^2}{N \operatorname{var} x} \tag{1}$$

Now, to complete the expansion, we must evaluate the second derivative at a_1:

$$\frac{d^2\Phi}{da^2} = -3\frac{1-a_1^2}{(1+a_1^2)^2} - \frac{N}{\sigma^2}\operatorname{var}\, x = -\frac{1}{\sigma_1^2} \tag{2}$$

$$\Phi(a) \approx \Phi(a_1) + \frac{1}{2}\frac{d^2\Phi}{da^2}\bigg|_{a_1}(a-a_1)^2 = \Phi(a_1) - \frac{(a-a_1)^2}{2\sigma_1^2}$$

we find this by using equations (1) and (2)

Now we complete the evaluation of the integral

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} \frac{da}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)\right] \\
&\quad = \int_{-\infty}^{+\infty} \exp\left[\Phi(a)\right] da \\
&\quad \approx \int_{-\infty}^{+\infty} \exp\left[\Phi(a_1) - \frac{\left(a-a_1\right)^2}{2\sigma_1^2}\right] da = \sqrt{2\pi\sigma_1^2}\exp\left[\Phi(a_1)\right]
\end{aligned}
$$

and finally, we find the posterior distribution:

$$
\begin{aligned}
p\left(a,b \mid \mathbf{y},\mathbf{x},\sigma\right) &\propto \frac{1}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right] \\
&\approx \exp\left[-\Phi(a_1) - \frac{\left(a-a_1\right)^2}{2\sigma_1^2}\right] \exp\left[-\frac{N}{2\sigma^2}\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right)\right)^2\right]
\end{aligned}
$$

From the posterior

$$p\left(a,b\,|\,\mathbf{y},\mathbf{x},\sigma\right) \propto \frac{1}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]$$

$$\approx \exp\left[-\Phi\left(a_1\right) - \frac{\left(a-a_1\right)^2}{2\sigma_1^2}\right] \exp\left[-\frac{N}{2\sigma^2}\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right)\right)^2\right]$$

we see that

$$\langle a \rangle = a_1; \quad \operatorname{var} a = \sigma_1^2;$$

$$\langle b \rangle = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right); \quad \operatorname{var} b = \frac{\sigma^2}{N}$$