THE BAYES-LAPLACE INFERENTIAL PARADIGM

Luca La Rocca
http://www-dimat.unipv.it/luca

April, 2009
Revised October 26, 2012

Outline

Proportion estimation

Model choice

Proportion estimation

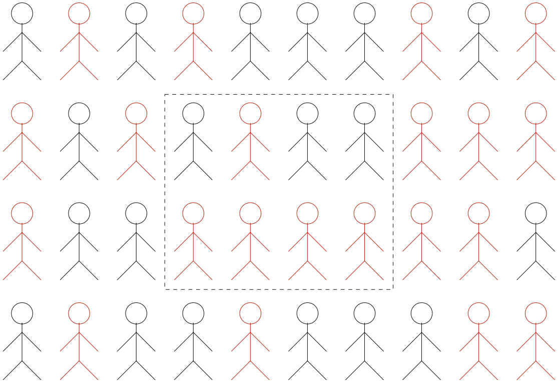
- Frequentist inference
- Bayesian inference

Model choice

We introduce the Bayes-Laplace inferential paradigm in the context of a typical statistical scenario: suppose that

- we have a population of N individuals who may agree or not on a given proposal, with N too large (in practice) to ask everyone
- we sample (at random) n individuals from the population, and record that pn of them agree on the proposal
- based on the sample proportion p, we would like to draw some conclusions (make inference) on the population proportion ψ of individuals who agree on the proposal

Toy example: how many red people? ($N = 40$, $n = 8$, $\psi = 0.525$, $p = 0.625$)



Real-life example: based on a survey by “Observa - Science in Society”, joint with “Tutto Scienze - La Stampa”, “SuperQuark” and “Quark”, published on “Tutto Scienze - La Stampa” on November 30, 2005, we know that

> *a proportion* $p = 52.7\%$ *of* $n = 1011$ *interviewed people, sampled from the Italian population aged 15 or more, would accept to “give up their car or motorbike and use bike and public transport” to reduce traffic pollution*

In order to make a decision, we could like to answer the question

> *is the proportion* ψ *of Italians aged 15 or more who would accept the sacrifice greater than 50%, i.e., are they a majority?*

Notice that N is not given, but we know that $n \ll N$.

Since the sample is taken at random, in order to avoid selection bias,

$$Y_i = \begin{cases} 1 & \text{if the } i^{th} \text{ interviewed individual agrees} \\ 0 & \text{if the } i^{th} \text{ interviewed individual disagrees} \end{cases}$$

for $i = 1, \ldots, n$ are random variables, observed when the survey is carried out; if all individuals in the population have the same probability, $\frac{1}{N}$, of being interviewed, that is, under simple random sampling, classical probability gives

$$\mathbb{P}(Y_i = 1|\psi) = \psi$$

as there are ψN individuals out of N who agree on the proposal.

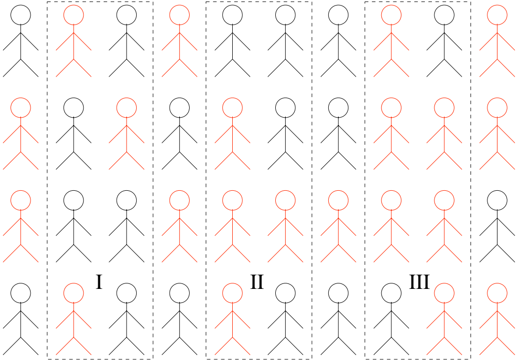
In the following, for illustrative purposes, we shall pretend that the sample in the pollution example be simple; in fact it is stratified.

We shall also pretend that people are interviewed one at a time, forgetting their names immediately after the interview, so that having been already interviewed does not change your probability of being interviewed next: this is known as (simple random) sampling with replacement; under this sampling scheme $Y_1, \dots, Y_n$ are conditionally independent given ψ and this gives a binomial distribution

$$\mathbb{P}(S = s|\psi) = \frac{n!}{s!(n-s)!}\psi^s(1-\psi)^{n-s}, \qquad s = 0, \dots, n$$

for the sample count $S = Y_1 + \cdots + Y_n$.

Sampling variability in the toy example ($S_I = 3$, $S_{II} = 4$, $S_{III} = 6$):



In practice, of course, there is no reason to interview the same person twice, i.e., we sample without replacement; however, it turns out that for $n \ll N$ the above binomial distribution is a good approximation to the exact sampling distribution (a hypergeometric distribution).

A well-established frequentist procedure to make inference on ψ—see for instance Agresti & Finlay (1997)—constructs an interval containing the sample count S with high probability, conditionally on $\psi\ldots$

Proportion estimation

- Frequentist inference
- Bayesian inference

Model choice

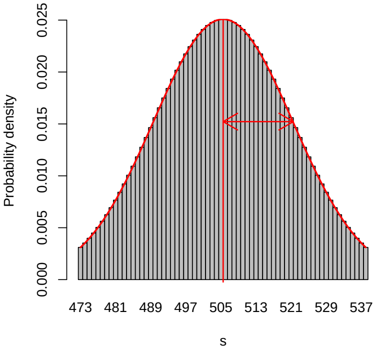
Proportion estimation

Frequentist inference for a binomial proportion, ψ, is usually based on a normal approximation to the binomial likelihood

$$\mathbb{P}(S = s|\psi) \simeq \int_{s-\frac{1}{2}}^{s+\frac{1}{2}} \frac{1}{\sqrt{2\pi n\psi(1-\psi)}} \exp\left\{-\frac{(x-n\psi)^2}{2n\psi(1-\psi)}\right\} dx$$

preserving the expected value and the standard deviation of S given ψ:

$$\begin{aligned} \mathbb{E}[S|\psi] &= \sum_{s=0}^{n} s\,\mathbb{P}(S = s|\psi) &= n\psi \\ \mathit{sd}(S|\psi) &= \sqrt{\mathbb{E}[(S-n\psi)^2|\psi]} &= \sqrt{n\psi(1-\psi)} \end{aligned}$$



$$n = 1011 \quad \psi = 0.5 \quad \Rightarrow \quad n\psi \simeq 505.5 \quad \sqrt{n\psi(1-\psi)} \simeq 15.9$$

Whatever the value of ψ, we have

$$\mathbb{P}\left(n\psi - 2 \cdot \mathit{sd}\left(S|\psi\right) \leq S \leq n\psi + 2 \cdot \mathit{sd}\left(S|\psi\right) \mid \psi\right) \simeq 95\%$$

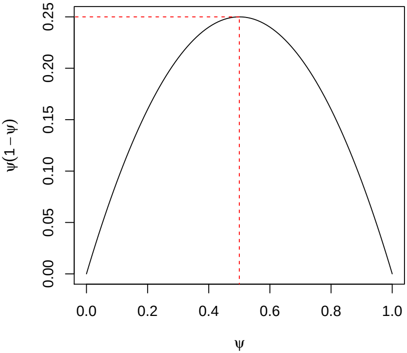
where

- 95% is a conventionally chosen confidence level
- 2 is the confidence coefficient for the chosen confidence level, obtained from a table of the normal distribution

Hence, by inequality manipulation, we find

$$\mathbb{P}\left(\frac{S}{n} - 2\frac{\mathit{sd}\left(S|\psi\right)}{n} \leq \psi \leq \frac{S}{n} + 2\frac{\mathit{sd}\left(S|\psi\right)}{n} \,\middle|\, \psi \right) \simeq 95\%$$

Note that $sd(S|\psi) = \sqrt{n\psi(1-\psi)}$ reaches its maximum, $\sqrt{n/4}$, for $\psi = 50\%$:



Thus, some simple algebra gives us

$$\mathbb{P}\left(\frac{S}{n} - \sqrt{\frac{1}{n}} \leq \psi \leq \frac{S}{n} + \sqrt{\frac{1}{n}} \,\middle|\, \psi\right) \gtrsim 95\%$$

that is, a random interval

- including ψ with high probability (whatever its actual value)
- whose endpoints only depend on the statistic S (and not on the unknown parameter ψ)

We call it a confidence interval of level (greater than) 95%.

Once the survey data are available, we observe $S/n = p$ and claim that

we are confident at level (greater than) 95% that

$$p - \frac{1}{\sqrt{n}} \leq \psi \leq p + \frac{1}{\sqrt{n}}$$

which in the pollution example translates to

$$49.6\% \leq \psi \leq 55.8\%$$

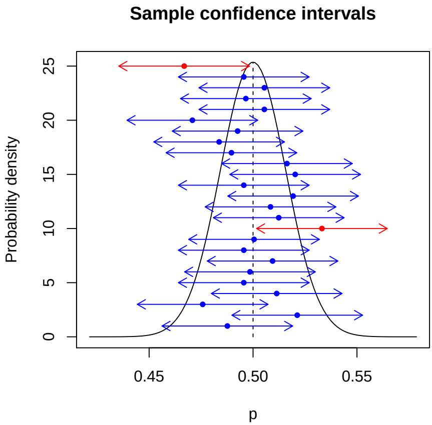
so that the null hypothesis $\psi \leq 50\%$ *is not ruled out.*

What does it mean?

It means that, unless we have been unlucky when selecting the sample, which happens (less than) once every twenty samples (in the long run) it holds that 49.6% $\leq \psi \leq$ 55.8%.

It does not mean that 49.6% $\leq \psi \leq$ 55.8% with probability 95%, as the unknown parameter is not dealt with as a random quantity.

The following figure (drawn letting $\psi = 0.5$) shows that most sample confidence intervals (the blue ones) contain the unknown population proportion; yet, if we get a bad (red) one, we are simply wrong...


Sample confidence intervals
Probability density
0
5
10
15
20
25
0.45
0.50
0.55
p

If you feel that confidence intervals

- do not have the meaning you would like them to have, that is, a meaning which can be interpreted directly
- have too much to do with the many samples you could have picked, and too little with the single one you got
- do not represent a valid state of knowledge, as they cannot be used to bet on the parameter value

you might be interested in some 18th century work by Thomas Bayes and Pierre Simon Laplace. . .

Proportion estimation

Frequentist inference

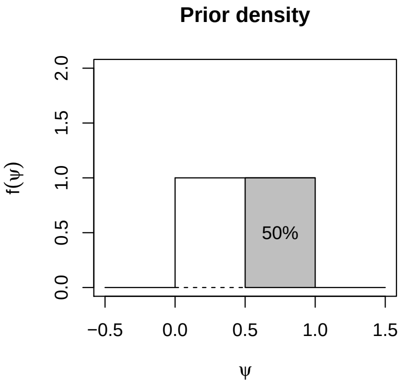
Bayesian inference

Model choice

Bayes (1763) studied the following problem, whose description here is adapted from Robert (2001):

> *A billiard ball is rolled on a line of length one, with a uniform probability of stopping everywhere. It stops at ψ.*
> *A second ball is rolled n times, under the same assumptions.*
> *Let S be the number of times it stopped on the left of the first ball. Given S, what inference can we make on ψ?*

The distribution of S given ψ is clearly the same as in the pollution example, though S and ψ have different meanings; here, however, we also know that $\mathbb{P}(a \leq \psi \leq b) = b - a$ for $0 \leq a \leq b \leq 1 \ldots$



$$f(\psi) = \begin{cases} 1 & \text{if } 0 \leq \psi \leq 1 \\ 0 & \text{if } \psi < 0 \text{ or } \psi > 1 \end{cases}$$

…armed with this uniform prior distribution for the unknown parameter, we have a full probabilistic model (for data and parameter) and we can make inference by conditioning on data.

The posterior density of ψ given the observed event $\{S = s\}$ can be found by applying a version of Bayes formula:

$$f(\psi|S = s) \quad \propto \quad \mathbb{P}(S = s|\psi)f(\psi) \quad \propto \quad \psi^s(1 - \psi)^{n-s}$$

with normalization constant given by the constraint

$$\int_0^1 f(\psi|S = s)\, d\psi \quad = \quad 1$$

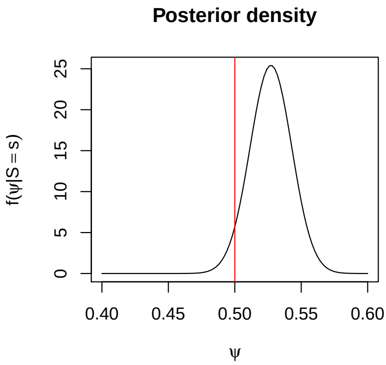
It is known from mathematical analysis that

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx, \qquad a > 0 \,\&\, b > 0$$

defines the beta function, and that $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,dx$, $a > 0$ (the well known gamma function such that $\Gamma(n+1) = n!$ for all positive integer n); hence, we find

$$f(\psi|S=s) = \frac{\psi^{(s+1)-1}(1-\psi)^{(n-s+1)-1}}{B(s+1, n-s+1)}, \qquad 0 \leq \psi \leq 1$$

called a beta density with parameters $s+1$ and $n-s+1\ldots$



$$n = 1011 \quad p = 0.527 \quad \Rightarrow \quad s = 533 \quad n - s = 478$$

…having switched to the pollution example.

Here we could like to compute the posterior probability of “majority”:

$$\mathbb{P}(\psi > 50\% | S = s) \quad = \quad \int_{\frac{1}{2}}^{1} f(\psi | S = s)\, d\psi$$

where ψ is (ab)used both for the parameter and its possible values.

Using our favourite statistical software, we find

$$\mathbb{P}(\psi > 0.5 | S = 533, n = 1011) \quad \simeq \quad 95.8\%$$

against the initial 50% (uniform distribution).

Alternatively, we may like to construct a posterior 95% credible interval for ψ by finding $\psi_\ell \simeq 0.496$ and $\psi_u \simeq 0.558$ such that

$$\mathbb{P}(\psi < \psi_\ell | S = s) = \mathbb{P}(\psi > \psi_u | S = s) = 2.5\%$$

and consequently $\mathbb{P}(\psi_\ell \leq \psi \leq \psi_u | S = s) \simeq 95\%$.

It is immediate to check that, for the data at hand, the credible interval gives us the same numerical answer as the confidence interval…

…but with a direct interpretation: $49.6\% \leq \psi \leq 55.8\%$ with probability 95%; it is not uncommon for the Bayesian approach to set a frequentist answer on firmer ground.

From an instrumental point of view, a prior distribution can be seen as a tool to derive “good” frequentist procedures; see Robert (2001).

Here, however, we insist on prior distributions representing prior knowledge on unknown parameters:

> *why, then, a uniform prior distribution for ψ?*

Laplace suggested that, in lack of specific prior information, this was motivated by the fact that no value of ψ should be preferred to the other ones; this was later referred to as the principle of insufficient reason.

Unfortunately, this argument is not invariant to reparameterization. . .

Proportion estimation

. . . consider, for instance, a car having covered 6 Km at unknown speed; if all we know is that the speed was between 60 Km/h and 120 Km/h, the principle of insufficient reason gives equal probability to the events

$$\begin{aligned} A &= \text{"speed between 60 Km/h and 90 Km/h"} \\ B &= \text{"speed between 90 Km/h and 120 Km/h"} \end{aligned}$$

which, in terms of travel time, translate to

$$\begin{aligned} A &= \text{"travel time between 4 min and 6 min"} \\ B &= \text{"travel time between 3 min and 4 min"} \end{aligned}$$

so that the resulting travel time distribution is not uniform, though we only know that the travel time was between 3 min and 6 min.

So...what prior for ψ? It is useful to have a family of prior distributions indexed by a few hyperparameters which can be tuned to account for prior information; for an unknown proportion, a convenient choice is

$$f_{a,b}(\psi) = \frac{\psi^{a-1}(1-\psi)^{b-1}}{B(a,b)}$$

for $a > 0$ and $b > 0$, including the uniform distribution as a special case ($a = b = 1$). This family is a convenient choice because it is conjugate, with respect to the binomial likelihood, i.e., it includes all posteriors

$$f_{a,b}(\psi | S = s) = \frac{\psi^{(s+a)-1}(1-\psi)^{(n-s+b)-1}}{B(s+a, n-s+b)}$$

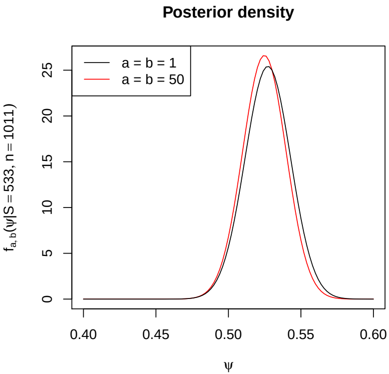
How can we choose *a* and *b*? Inspection of the formula for the posterior density shows that

- *a* has the meaning of a prior count of people who agree
- *b* has the meaning of a prior count of people who disagree
- $a + b$ has the meaning of a prior sample size

Typical choices are

- $a = b = 1$ (uniform)
- $a = b = \frac{1}{2}$ (Jeffreys)
- $a = b = 0$ (improper)

the latter meaning that we base inference on a limit posterior.



Hopefully, for large samples ($a + b \ll n$) the actual choice of a and b will have limited impact on the posterior distribution: posterior consistency.

One of the advantages of the Bayesian approach is that it provides us with the predictive distribution of the next observation given the data:

$$\begin{aligned}
\mathbb{P}(Y_{n+1} = 1 | S = s) &= \mathbb{E}[\psi | S = s] \\
&= \frac{a+s}{a+b+s+(n-s)} \\
&= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{s}{n} \\
&= \frac{a+b}{a+b+n}\mathbb{E}[\psi] + \frac{n}{a+b+n}p
\end{aligned}$$

When n is zero, we accept bets based on $\mathbb{E}[\psi]$ (prior expected value); as $n \to \infty$, we end up accepting bets based on p (sample proportion).

Proportion estimation

Model choice

Following Jefferys & Berger (1992) we travel back in time to 1920:

- Newtonian theory had been successful in explaining most of the motions in the solar system, but for an observed residual motion of Mercury's perihelion (the point in its orbit closest to the Sun) of approximately $y = 41.6$ seconds of arc per century
- the American astronomer Simon Newcomb had suggested, in 1895, that the exponent of Newton's law of gravity be $2+\epsilon$, instead of 2, leading to an unknown residual motion μ
- Einstein's theory of general relativity, announced in 1915, implied an exact value of $\mu_0 = 42.9$ seconds of arc per century

Two competing explanations for the same evidence... who is right?

Since the observed residual motion $y = 41.6''$ is affected by measurement error, we assume (central limit theorem) that it comes from a normal sampling distribution with density

$$f_\sigma(y|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \qquad y \in \mathbb{R}$$

where μ is the "true" residual motion, and $\sigma = 2.0''$ is the (inverse of) measurement precision. Our knowledge about μ is

- if Einstein is right (the null model, denoted by $\mathcal{M}_0$) then $\mu = \mu_0 = 42.9''$ (degenerate prior distribution)
- if Newcomb is right (the alternative model, denoted by $\mathcal{M}_1$) then μ deserves a convenient (non-degenerate) prior distribution

As the Newtonian theory is well-established, our prior distribution under Newcomb's model should favour values close to $\mu = 0$; in particular, we should like to use a prior density $f(\mu|\mathcal{M}_1)$, $\mu \in \mathbb{R}$, such that

(i) $f(\mu|\mathcal{M}_1)$ is maximum for $\mu = 0$

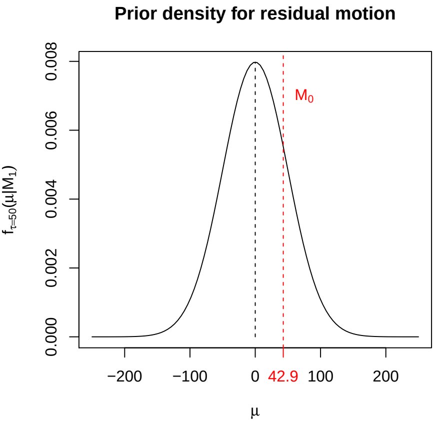
(ii) $f(\mu|\mathcal{M}_1)$ decreases as μ moves away from zero

(ii) $f(\mu|\mathcal{M}_1)$ is equal to $f(-\mu|\mathcal{M}_1)$, i.e., it is symmetrical

A convenient (conjugate to the normal likelihood) choice is

$$f_\tau(\mu|\mathcal{M}_1) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{\mu^2}{2\tau^2}\right\}, \qquad \mu \in \mathbb{R}$$

where $\tau = 50''$ by comparison with the orbits of other inner planets.


Prior density for residual motion
$f_{\tau=50}(\mu|M_1)$
0.000
0.002
0.004
0.006
0.008
−200
−100
0
42.9
100
200
μ
M_0

We are now in a position to compute the marginal density of the observed residual motion (evidence) both under $\mathcal{M}_0$ and $\mathcal{M}_1$:

$$
\begin{aligned}
f(y|\mathcal{M}_0) &= f_\sigma(y|\mu_0) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu_0)^2}{2\sigma^2}\right\} \\
f(y|\mathcal{M}_1) &= \int_{-\infty}^{+\infty} f_\sigma(y|\mu) f_\tau(\mu|\mathcal{M}_1)\, d\mu \\
&= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left\{-\frac{y^2}{2(\sigma^2+\tau^2)}\right\}
\end{aligned}
$$

By plugging in the values $y = 41.6$, $\mu_0 = 42.9$, $\sigma = 2.0$ and $\tau = 50$, we obtain $f(y|\mathcal{M}_0) = 0.1614862$ and $f(y|\mathcal{M}_1) = 0.00564311\ldots$

…thus $\mathcal{M}_0$ has a much greater (marginal) likelihood than $\mathcal{M}_1$ with respect to our data.

In order to conclude our Bayesian analysis, we need to introduce some prior model probabilities: for instance, if *a priori* we give the same credit to Einstein and Newcomb, we shall take

$$\mathbb{P}(\mathcal{M}_0) = \mathbb{P}(\mathcal{M}_1) = 0.5$$

Then, a version of Bayes theorem…

…will give us the corresponding posterior model probabilities:

$$\begin{aligned}\mathbb{P}(\mathcal{M}_0|y) &\propto f(y|\mathcal{M}_0)\mathbb{P}(\mathcal{M}_0)\\ \mathbb{P}(\mathcal{M}_1|y) &\propto f(y|\mathcal{M}_1)\mathbb{P}(\mathcal{M}_1)\end{aligned}$$

with normalization constant given by $\mathbb{P}(\mathcal{M}_0|y) + \mathbb{P}(\mathcal{M}_1|y) = 1$.

In this way, we find $\mathbb{P}(\mathcal{M}_0|y) \simeq 96.6\%$ and $\mathbb{P}(\mathcal{M}_1|y) \simeq 3.4\%$: we would pay slightly more than 28 : 1 for a bet on Newcomb.

Since $\mathbb{P}(\mathcal{M}_0) = \mathbb{P}(\mathcal{M}_1)$ the posterior probability of each model is proportional to its (marginal) likelihood; this is not always the case, but…

. . . it is always possible to write

$$\frac{\mathbb{P}(\mathcal{M}_1|y)}{\mathbb{P}(\mathcal{M}_0|y)} = \frac{f(y|\mathcal{M}_1)}{f(y|\mathcal{M}_0)} \frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(\mathcal{M}_0)}$$

so that the effect of prior odds on posterior odds is separated from the weight of evidence for $\mathcal{M}_1$ against $\mathcal{M}_0$, the latter being measured by the Bayes factor

$$BF = \frac{f(y|\mathcal{M}_1)}{f(y|\mathcal{M}_0)}$$

which in our example was about 0.035 (slightly less than 1/28).

Notice, however, that the Bayes factor still depends on the prior distributions under $\mathcal{M}_0$ and $\mathcal{M}_1$: in our case the prior under $\mathcal{M}_0$ has no tuning hyperparameter, whereas the prior under $\mathcal{M}_1$ depends on τ...

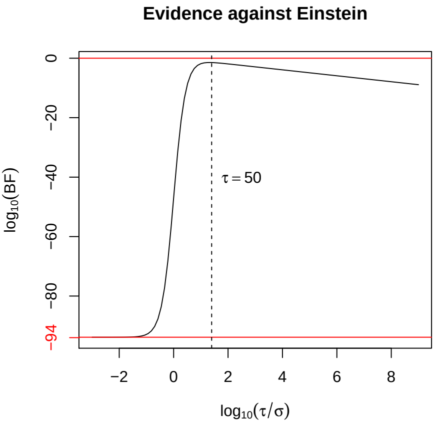
...and the choice of τ is critical:

$$\ln BF = \frac{(y-\mu_0)^2}{2\sigma^2} - \frac{y^2}{2\sigma^2(1+\tau^2/\sigma^2)} - \frac{1}{2}\ln(1+\tau^2/\sigma^2)$$

The next figure plots

$$\log_{10}(BF) = \frac{\ln BF}{\ln 10}$$

as a function of τ...


Evidence against Einstein
log₁₀(BF)
0
−20
−40
−60
−80
−94
τ = 50
−2
0
2
4
6
8
log₁₀(τ/σ)

Hence, as a function of τ, the BF behaves as follows:

$$\log_{10}(BF) \simeq \frac{(y-\mu_0)^2 - y^2}{2\sigma^2 \ln 10} \simeq -94$$

for $\tau \ll \sigma$, that is, Einstein is better than Newton, and for $\tau \gg \sigma$

$$\log_{10}(BF) \simeq \frac{(y-\mu_0)^2}{2\sigma^2 \ln 10} - \log_{10}(\tau/\sigma) \simeq 0.1 - \log_{10}(\tau/\sigma)$$

i.e., Einstein is “definitely” better than Newcomb ($BF \to 0$, as $\tau \to \infty$).

Note the difference between the posterior distribution of ψ, converging to a beta distribution with parameters s and $n-s$, as a and b tend to zero, and the Bayes factor, having no useful limit value as τ tends to infinity. . .

...quite remarkably, however, the Bayes factor has a maximum, which turns out to be very close to the actual value we got:

$$\max_{\tau} BF \simeq 0.036$$

Thus, with the data at hand, and with a zero-mean normal prior for μ under $\mathcal{M}_1$, evidence will always be in favour of $\mathcal{M}_0$; Jefferys & Berger (1992) show that all priors satisfying (i) to (iii) give a BF below 0.067.

This behaviour of BF implements Ockham's razor, the principle stated by the English theologian William of Ockham (circa 1290–circa 1349) that an explanation should not be more complicated than necessary:

PLURALITAS NON EST PONENDA SINE NECESSITATE

If we choose $\mathcal{M}_0$ and discard $\mathcal{M}_1$, the predictive density of the residual motion we shall measure in the next century will be

$$f(y_\star|y,\mathcal{M}_0) \quad = \quad f_\sigma(y_\star|\mu_0), \qquad y_\star \in \mathbb{R}$$

as we learn nothing on μ_0 from y—but we should update σ so as to account for technological progress—whereas choosing $\mathcal{M}_1$ it would be

$$\begin{aligned} f(y_\star|y,\mathcal{M}_1) &= \int_{-\infty}^{+\infty} f_\sigma(y_\star|\mu) f(\mu|y,\mathcal{M}_1)\,d\mu \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau_\star^2)}} \exp\left\{-\frac{y_\star^2}{2(\sigma^2+\tau_\star^2)}\right\} \end{aligned}$$

where $\tau_\star^2 = (\sigma^{-2}+\tau^{-2})^{-1}$ is the posterior variance of μ under $\mathcal{M}_1$.

When posterior model probabilities are close to each other, we may not like the idea of choosing a model and discarding the other one(s). . .

. . . the Bayesian approach does not force as to do so, as we may instead resort to model averaging:

$$f(y_\star|y) = f(y_\star|y,\mathcal{M}_0)\mathbb{P}(\mathcal{M}_0|y) + f(y_\star|y,\mathcal{M}_1)\mathbb{P}(\mathcal{M}_1|y)$$

The above marginal predictive density incorporates in the prediction both the uncertainty on the unknown parameter(s) and the uncertainty on the model; generalization to more than two models is straightforward, though prior specification for many models is challenging.

References

Agresti, A. & Finlay, B. (1997).
Statistical Methods for the Social Sciences. Third Edition.
Prentice-Hall.

Bayes, T. (1763).
An essay towards solving a problem in the doctrine of chances.
Phil. Trans. Roy. Soc. **53**, 370–418.

Jefferys, W. & Berger, J.O. (1992).
Ockham's razor and Bayesian analysis.
American Scientist **80**, 64–72.

Robert C. P. (2001).
The Bayesian Choice. Second Edition.
Springer-Verlag.