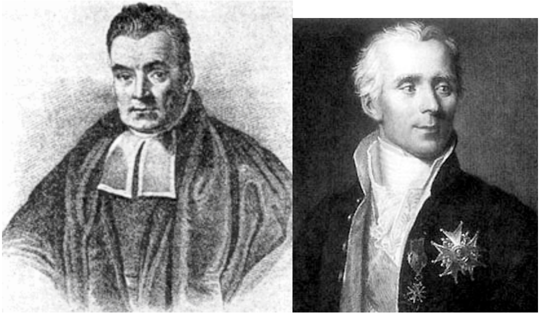
Bayes' Theorem

Bayesian Models for Ecologists

Tom Hobbs

June 4, 2024

Middlebury Funga THE UNIVERSITY OF TEXAS AT AUSTIN Colorado State University NSF DEB 2042028



See
https://www.tcbegley.com/blog/posts/bayesian-billiards

Laplace's theory of "inverse probability"

$$P(C|E) = \frac{P(E|C)P_{\text{prior}}(C)}{\sum(E|C')P_{\text{prior}}(C')} \tag{1}$$

ferent. I propose to determine the probability of the causes of events, a question which has not been given due consideration before, but which deserves even more to be studied, for it is principally from this point of view that the science of chances can be useful in civil life.

Translated from the original French by S. M. Stigler, University of Chicago. Originally published as "Mémoire sur la probabilité des causes par les évènemens," par M. de la Place, Professeur à l'École royal Militaire, in *Mémoires de Mathématique et de Physique, Presentés à l'Académie Royale des Sciences, par divers Savans & lûs dans ses Assemblées, Tome Sixième* (1774) 621–656. Reprinted in Laplace's Oeuvres complètes **8** 27–65.

THÉORIE

ANALYTIQUE

DES PROBABILITES;

Par M. LE COMTE LAPLACE,

Chancelier du Sénat-Conservateur, Grand-Officier de la Légion d'Honneur; Membre de l'Institut impérial et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, de Hollande, d'Italie, etc.

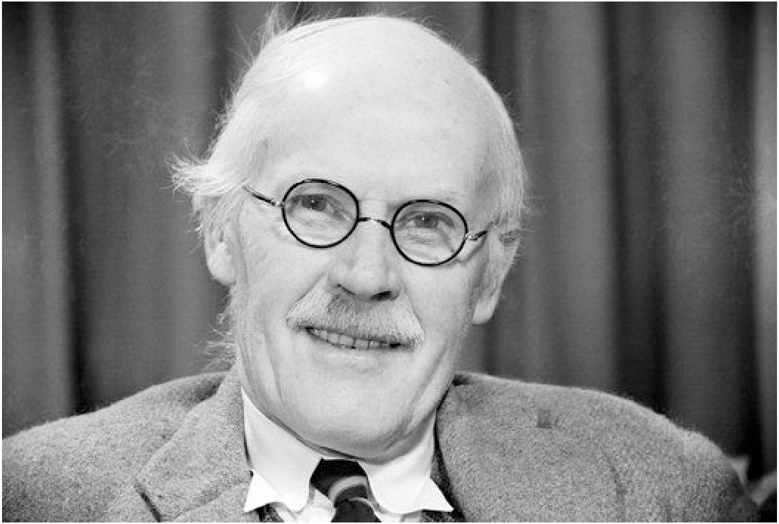
PARIS,

Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, n° 57.

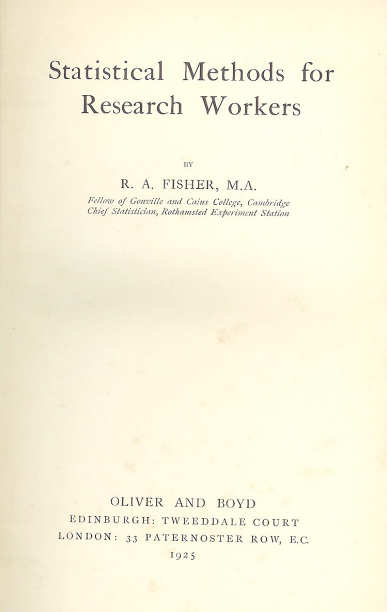
1812.

"Laplace's principle being dead, it should be decently buried out of sight, and not embalmed in text-books and examination papers... The indiscretions of great men should be quietly allowed to be forgotten."

George Chrystal 1891



The frequentist view



Statistical Methods for Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

"My personal conviction is that the theory of inverse probability is founded upon an error and must be wholly rejected"
R. A. Fischer

"There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected the law outright."
Harold Jefferys

"The p-value is almost nothing sensible you can think of. I tell students to give up trying."
Stephen Goodman

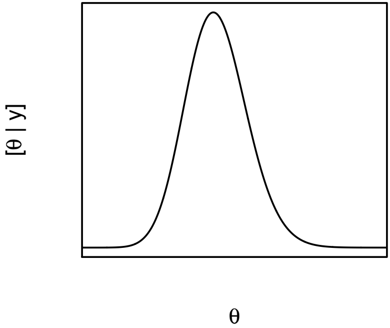
What is the collective noun for a group of statisticians?


ENIGMA

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85:398-409.

Bayesian inference

All unobserved quantities are treated as *random variables*.



Random variables

All unobserved quantities are treated in exactly the same way.

- model parameters
- latent states
- missing data
- predictions and forecasts
- observations (before they are observed)

Exercise

- Assume we have two, jointly distributed random variables, θ and y. The random variable θ represents unobserved quantities of interest. The random variable y represents observations, which become fixed *after* they are observed.
- Derive Bayes' Theorem

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]} \tag{2}$$

using your knowledge of the laws of probability, particularly the definition of conditional probability.

Derivation

Recall the definition of conditional probability

$$[\theta|y] = \frac{[\theta, y]}{[y]} \tag{3}$$

$$[y|\theta] = \frac{[\theta, y]}{[\theta]}. \tag{4}$$

Solving 4 for $[\theta, y]$

$$[\theta, y] = [y|\theta]\,[\theta]. \tag{5}$$

Substituting the right hand side of 5 for $[\theta, y]$ in 3 we obtain Bayes' Theorem

$$\boxed{[\theta|y] = \frac{[y|\theta]\,[\theta]}{[y]}.} \tag{6}$$

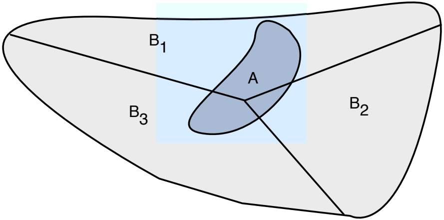
We will often make use of the equivalent equation

$$[\theta|y] = \frac{[y, \theta]}{[y]} \tag{7}$$

as a starting point for developing hierarchical models by factoring $[y, \theta]$ into ecologically sensible components that can be treated in MCMC as univariate distributions. More about that soon.

What is $[y]$?

Recall the law of total probability for discrete random variables



$$[A] = \sum_i [A \mid B_i]\,[B_i]\,. \tag{8}$$

and for continuous random variables

$$[A] = \int_B [A|B]\,[B]\,dB. \tag{9}$$

What is $[y]$?

It follows that

$$[y] = \sum_{\theta_i \in \{\Theta\}} [y|\theta_i]\,[\theta_i] \text{ for discrete parameters} \tag{10}$$

$$[y] = \int_\theta [y|\theta]\,[\theta]\,d\theta \text{ for continuous parameters.} \tag{11}$$

Thus, Bayes theorem for discrete valued parameters is

$$[\theta|y] = \frac{[y|\theta_i]\,[\theta_i]}{\sum_{\theta_i \in \{\Theta\}} [y|\theta_i]\,[\theta_i]} \tag{12}$$

and for parameters that are continuous,

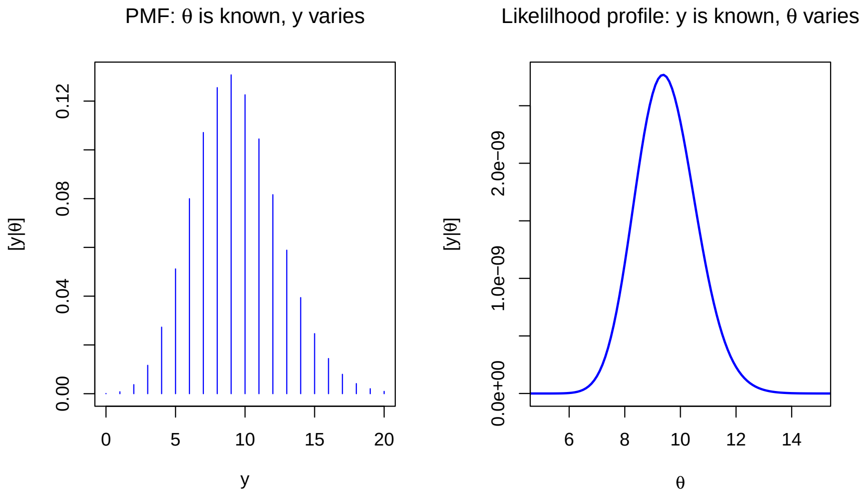
$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta}. \tag{13}$$

More about $[y]$

- $[y]$ is the marginal distribution of the data, a *distribution* before the data are observed and a *normalizing constant* after the data are observed.
- It is also called the *prior predictive distribution.* Why?
- Because $[y]$ is a constant after the data are observed,

$$
\begin{aligned}
[\theta|y] &\propto [y,\theta] && (14)\\
&\propto [y|\theta][\theta] && (15)
\end{aligned}
$$

More about $[y]$: Recall the PMF and likelihood profile

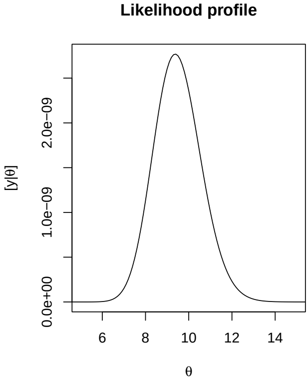


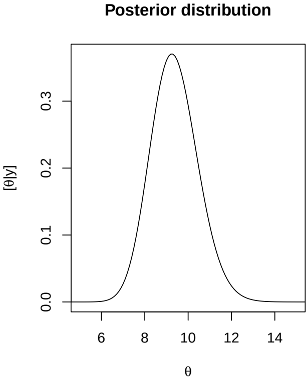
$[y]$ is critical to Bayes

$\mathbf{y} = (5, 10, 11, 12, 14, 9, 8, 6)'$

$\text{likelihood} = \prod_{i=1}^{8} \text{Poisson}(y_i|\theta)$

$\text{posterior} = \frac{\prod_{i=1}^{8} \text{Poisson}(y_i|\theta)\text{gamma}(\theta|.0001,.0001)}{[y]}$





Cut to example

Probability mass function $[y|\theta]$

θ is *known* to be $\frac{1}{2}$. Probability of number of whites conditional on three draws and $\theta = \frac{1}{2}$:

$y =$ Number of whites	$[y\|\theta]$
0	.125
1	.375
2	.375
3	.125
$\sum_{i=1}^{4} [y\|\theta_i] =$	1

New cans, switch to right board

Likelihood $[y|\theta]$

$\theta$is *unknown.* Probability of two whites on three draws conditional on θ_i

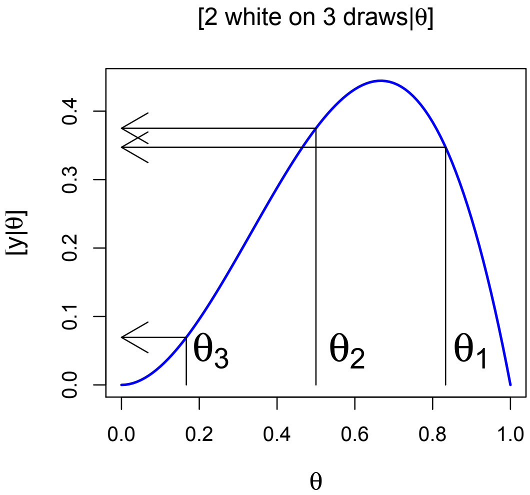
Hypothesis	Likelihood $[y|\theta_i]$
$\theta_1{=}{}^5/_6$	.347
$\theta_2{=}{}^1/_2$	.375
$\theta_3{=}{}^1/_6$	.069
$\sum_{i=1}^{3}[y|\theta_i] =$	.791

Posterior distribution $[\theta|y]$

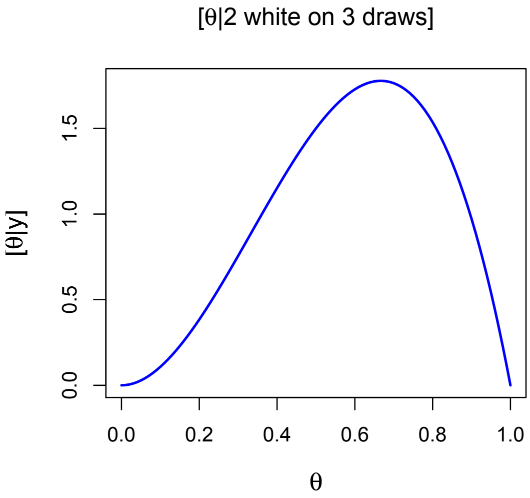
Probability of θ_i conditional on two whites on three draws

Hypothesis	Prior $[\theta_i]$	Likelihood $[y\|\theta_i]$	Joint $[y\|\theta_i]\,[\theta_i]$	Posterior $\frac{[y\|\theta_i][\theta_i]}{[y]} = [\theta_i\|y]$
$\theta_1 = 5/6$	0.333	0.347	0.115	0.439
$\theta_2 = 1/2$	0.333	0.375	0.125	0.474
$\theta_3 = 1/6$	0.333	0.069	0.023	0.087
		$[y] = \sum_{i=1}^{3} [y\|\theta_i]\,[\theta_i] =$	0.261	$\sum_{i=1}^{3} [\theta_i\|y] = 1$

Likelihood profile $[y|\theta]$



Posterior distribution $[\theta|y]$

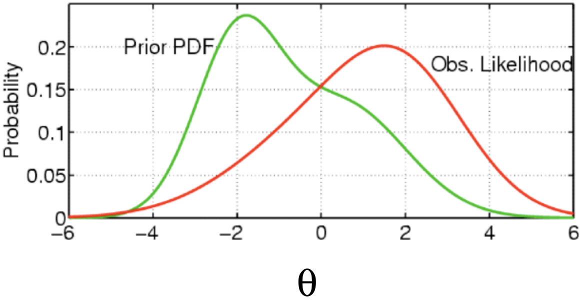


The components of Bayes theorem

$$\overbrace{[\theta|y]}^{\text{Posterior}} = \underbrace{\frac{\overbrace{[y|\theta]}^{\text{likelihood}} \; \overbrace{[\theta]}^{\text{prior}}}{\int_\theta [y|\theta]\,[\theta]\,d\theta}}_{} \tag{16}$$

$$\overbrace{[\theta|y]}^{\text{Posterior}} = \frac{\overbrace{[y|\theta]}^{\text{likelihood}} \; \overbrace{[\theta]}^{\text{prior}}}{\underbrace{\int_\theta [y|\theta]\,[\theta]\,d\theta}_{\text{marginal distribution of data}}}$$

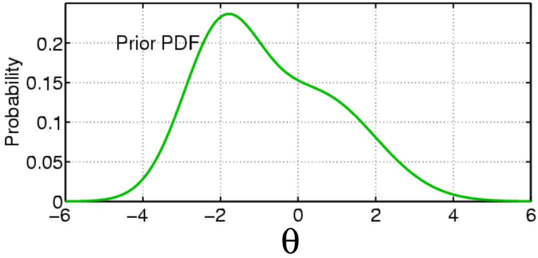
The components of Bayes Theorem



Courtesy of Chris Wikle, University of Missouri

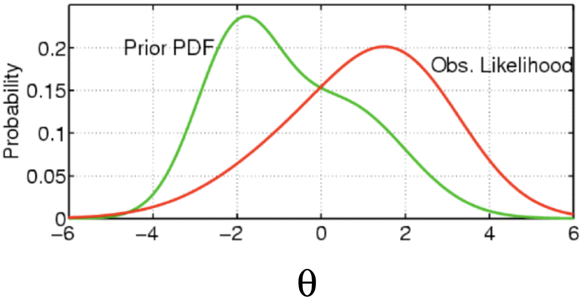
$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta} \tag{17}$$

The prior, $[\theta]$, can be informative or vague.



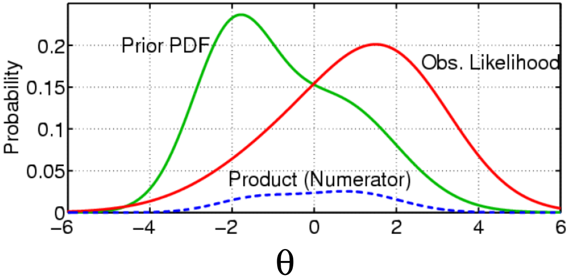
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta][\theta]\,d\theta} \tag{18}$$

The likelihood (a.k.a. data distribution, $[y|\theta]$)



$$[\theta|y] = \frac{[y,\theta]}{[y]} = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta} \tag{19}$$

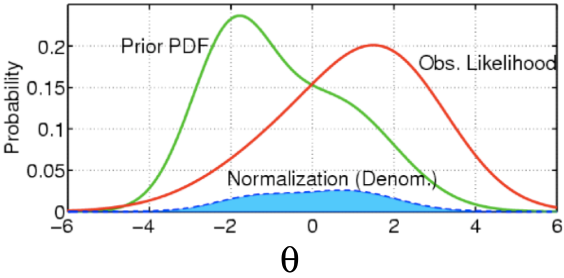
The product of the prior and the likelihood, $[y|\theta][\theta]$, the joint distribution of the parameters and the data, $[y,\theta]$.



What is the maximum likelihood estimate of θ?

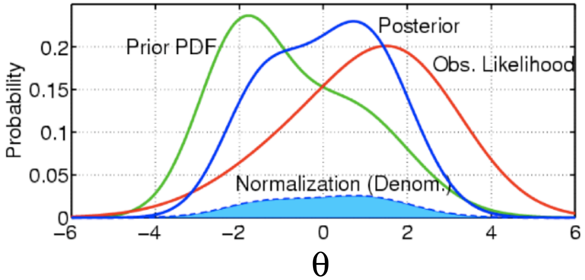
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta][\theta]\,d\theta} \tag{20}$$

The marginal distribution of the data (the denominator) is the area under the joint distribution.



What we are seeking: The posterior distribution, $[\theta|y]$.

$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta} \tag{21}$$



Note that we are dividing each point on the dashed line by the area under the dashed line to obtain a probability density function reflecting our prior and current knowledge about θ.

So what?

What does this enable you to do? Review factoring joint distributions:

Remember from the basic laws of probabilty that

$$p(z_1, z_2) = p(z_1 \mid z_2)\, p(z_2) = p(z_2 \mid z_1)\, p(z)_1$$

This genearlizes to:

$$\mathbf{z} = (z_{1,} z_2, \ldots z_n)$$

$$p(z_{1,} z_2, \ldots z_n) = p(z_n \mid z_{n-1 \ldots} z_1) \ldots . p(z_3 \mid z_2, z_1)\, p(z_2 \mid z_1)\, p(z_1)$$

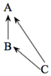
where the components z_i may be scalars or subvectors of $\mathbf{z}$ and the sequence of their conditioning is arbitrary. This equation can be simplified using knowledge of indepdence.

So what?

I

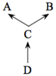
$$\Pr(A, B) = \Pr(A|B)\Pr(B)$$

II



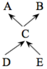
$$\Pr(A, B, C) = \Pr(A|B, C) \times \Pr(B|C)\Pr(C)$$

III



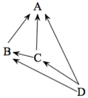
$$\Pr(A, B, C, D) = \Pr(A|C) \times \Pr(B|C)\Pr(C|D)\Pr(D)$$

IV



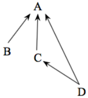
$$\Pr(A, B, C, D, E) = \Pr(A|C) \times \Pr(B|C)\Pr(C|D, E) \times \Pr(D)\Pr(E)$$

V



$$\Pr(A, B, C, D) = \Pr(A|B, C, D) \times \Pr(B|C, D) \times \Pr(C|D)\Pr(D)$$

VI



$$\Pr(A, B, C, D) = \Pr(A|B, C, D) \times \Pr(C|D) \times \Pr(B) Pr(D)$$

So what?

$$\overbrace{[\theta|y]}^{\text{Posterior}} = \frac{[y,\theta]}{[y]} = \frac{\overbrace{[y|\theta]}^{\text{likelihood}}\;\overbrace{[\theta]}^{\text{prior}}}{\underbrace{\int_\theta [y|\theta]\,[\theta]\,d\theta}_{\text{marginal}}} \tag{22}$$

Useful models will be more complex:

$$\underbrace{[\theta_1,\theta_2,\theta_{3,}...,\theta_n,\mathbf{z}_1,\mathbf{z}_2..\mathbf{z}_n|\mathbf{y}_1,\mathbf{y}_2]}_{\text{multiple parameters, latent states, data sets}} \propto \underbrace{[\theta_1,\theta_2,\theta_{3,}...,\theta_n,\mathbf{z}_1,\mathbf{z}_2..\mathbf{z}_n,\mathbf{y}_1,\mathbf{y}_2]}_{\text{factor into conidtional distributions}}$$

We use the rules of probability to factor complex joint distributions into a series of conditional distributions. We can then use the Markov chain Monte Carlo algorithm to escape the need for integrating the marginal data distribution, allowing us to find the marginal posterior distributions of all of the unobserved quantities. Which, of course, is where we started out. And where we are headed.