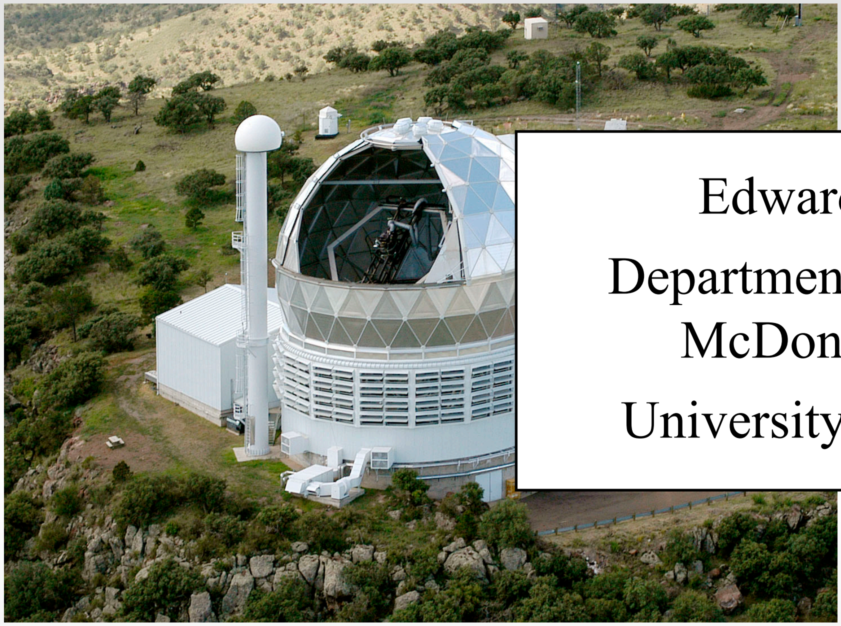
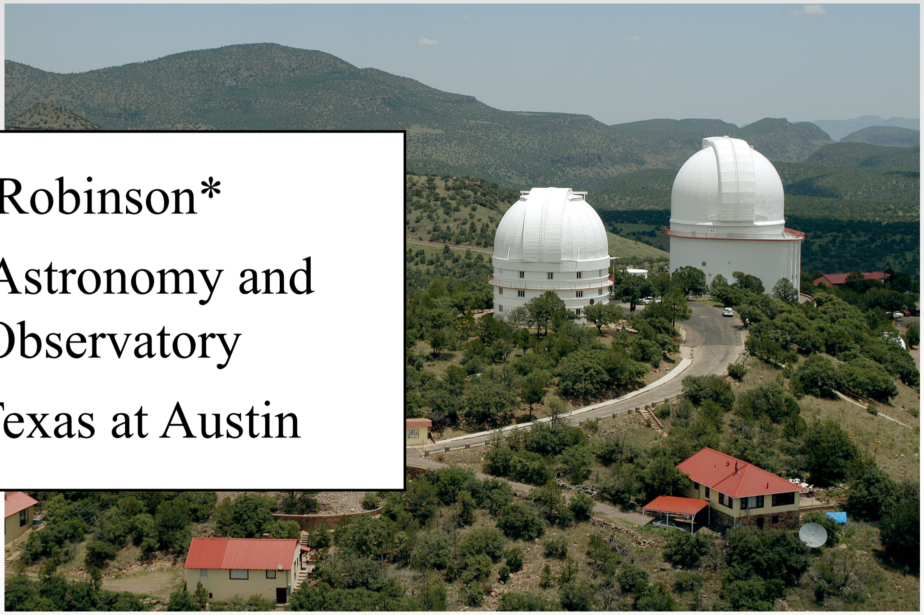
Introduction to Likelihood Statistics



Edward L. Robinson*

Department of Astronomy and
McDonald Observatory

University of Texas at Austin



1. The likelihood function.
2. Use of the likelihood function to model data.
3. Comparison to standard frequentist and Bayesean statistics.

*Look for: “Data Analysis for Scientists and Engineers” Princeton University Press, Sept 2016.

The Likelihood Function

- Let a probability distribution function for ξ have $m+1$ parameters a_j
$$f(\xi, a_0, a_1, \cdots, a_m) \;=\; f(\xi, \vec{a}),$$
The joint probability distribution for n samples of ξ is
$$f(\xi_1, \xi_2, \cdots, \xi_n, a_0, a_1, \cdots, a_m) \;=\; f(\vec{\xi}, \vec{a}).$$

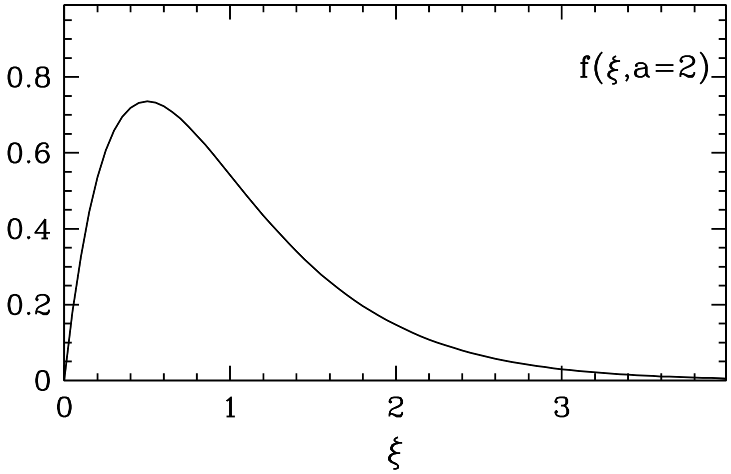
- Now make measurements. For each variable ξ_i there is a measured value x_i.

- To obtain the likelihood function $L(\vec{x}, \vec{a})$, replace each variable ξ_i with the numerical value of the corresponding data point x_i:
$$L(\vec{x}, \vec{a}) \;\equiv\; f(\vec{x}, \vec{a}) \;=\; f(x_1, x_2, \cdots, x_n, \vec{a}).$$
In the likelihood function the $\vec{x}$ are known and fixed, while the $\vec{a}$ are the variables.

A Simple Example

- Suppose the probability distribution for the data is

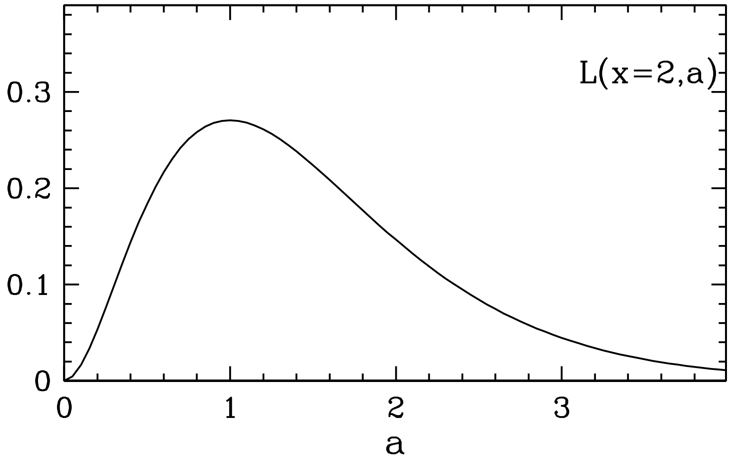
$$\mathbf{f(\xi, a) \;=\; a^2 \xi e^{-a\xi}}.$$



- Measure a single data point. It turns out to be $\mathbf{x = 2}$.

- The likelihood function is

$$\mathbf{L(x = 2, a) \;=\; 2a^2 e^{-2a}}.$$



A Somewhat More Realistic Example

Suppose we have n independent data points, each drawn from the same probability distribution

$$\mathbf{f}(\xi, \mathbf{a}) \;=\; \mathbf{a}^2 \xi \mathbf{e}^{-\mathbf{a}\xi}.$$

The joint probability distribution is

$$\begin{aligned} \mathbf{f}(\vec{\xi}, \mathbf{a}) &= \mathbf{f}(\xi_1, \mathbf{a})\mathbf{f}(\xi_2, \mathbf{a}) \cdots \mathbf{f}(\xi_n, \mathbf{a}) \\ &= \prod_{i=1}^{n} \mathbf{a}^2 \xi_i \mathbf{e}^{-\mathbf{a}\xi_i}. \end{aligned}$$

The data points are x_i. The resulting likelihood function is

$$\mathbf{L}(\vec{\mathbf{x}}, \mathbf{a}) \;=\; \prod_{i=1}^{n} \mathbf{a}^2 \mathbf{x}_i \mathbf{e}^{-\mathbf{a}\mathbf{x}_i}.$$

What Is the Likelihood Function? – 1

The likelihood function is not a probability distribution.

- It does not transform like a probability distribution.
- Normalization is not defined.

How do we deal with this?

Traditional approach: Use the **Likelihood Ratio**.

To compare the likelihood of two possible sets of parameters $\vec{a}_1$ and $\vec{a}_2$, construct the likelihood ratio:

$$\mathrm{LR} = \frac{L(\vec{x}, \vec{a}_1)}{L(\vec{x}, \vec{a}_2)} = \frac{f(\vec{x}, \vec{a}_1)}{f(\vec{x}, \vec{a}_2)}.$$

This is the ratio of the probabilities that data $\vec{x}$ would be produced by parameter values $\vec{a}_1$ and $\vec{a}_2$.

What Is the Likelihood Function? - 2

Compare *all* parameter values to a single set of fiducial parameter values $\vec{a}_0$. The likelihood ratio becomes

$$\mathrm{LR} \;=\; \frac{\mathrm{L}(\vec{x},\vec{a})}{\mathrm{L}(\vec{x},\vec{a}_0)} \;\propto\; \mathrm{L}(\vec{x},\vec{a}).$$

This likelihood ratio and therefore the likelihood function itself is proportional to the probability that the observed data $\vec{x}$ would be produced by parameter values $\vec{a}$.

What Is the Likelihood Function? - 3

An increasingly common and highly attractive approach (although it is unclear that everyone knows what they are doing):

Treat the likelihood function like a probability distribution!

- The quantity

  $$\mathrm{L}(\vec{\mathrm{x}}, \vec{\mathrm{a}})\, \mathrm{d}\vec{\mathrm{a}} \;=\; \mathrm{L}(\vec{\mathrm{x}}, \vec{\mathrm{a}})\, \mathrm{da}_0 \mathrm{da}_1 \mathrm{da}_2 \cdots \mathrm{da}_\mathrm{m}$$

  does transform like probability. This usage is entirely consistent with the standard definition of probability density.

- To normalize $\mathrm{L}(\vec{\mathrm{x}}, \vec{\mathrm{a}})$, define a multiplicative factor A such that

  $$1 \;=\; \mathrm{A} \int \mathrm{L}(\vec{\mathrm{x}}, \vec{\mathrm{a}})\, \mathrm{d}\vec{\mathrm{a}}.$$

  Then $\mathrm{AL}(\vec{\mathrm{x}}, \vec{\mathrm{a}})$ is normalized (but normalization never needed).

Likelihood Statistics

The likelihood function contains information about the new data.

(I am in the camp that says it contains *all* the new information.)

One can extract information from $\mathrm{L}(\vec{\mathrm{x}}, \vec{\mathrm{a}})$ in the same way one extracts information from an (un-normalized) probability distribution:

- calculate the mean, median, and mode of parameters.
- plot the likelihood and its marginal distributions.
- calculate variances and confidence intervals.
- Use it as a basis for χ^2 minimization!

But beware: One can usually get away with thinking of the likelihood function as the probability distribution for the parameters $\vec{\mathrm{a}}$, but this is not really correct. It is the probability that a specific set of parameters would yield the observed data.

The Maximum Likelihood Principle

The maximum likelihood principle is one way to extract information from the likelihood function. It says, in effect,

"Use the modal values of the parameters."

The Maximum Likelihood Principle

Given data points $\vec{x}$ drawn from a joint probability distribution whose functional form is known to be $f(\vec{\xi}, \vec{a})$, the best estimate of the parameters $\vec{a}$ are those that maximize the likelihood function $L(\vec{x}, \vec{a}) = f(\vec{x}, \vec{a})$.

Find the maximum values by setting

$$\left.\frac{\partial L}{\partial a_j}\right|_{\hat{a}} = 0.$$

These $m + 1$ equations are the "likelihood equations."

Maximum Likelihood Estimation

Recall our previous example: n independent data points x_i drawn from

$$f(\xi, a) \;=\; a^2 \xi e^{-a\xi}.$$

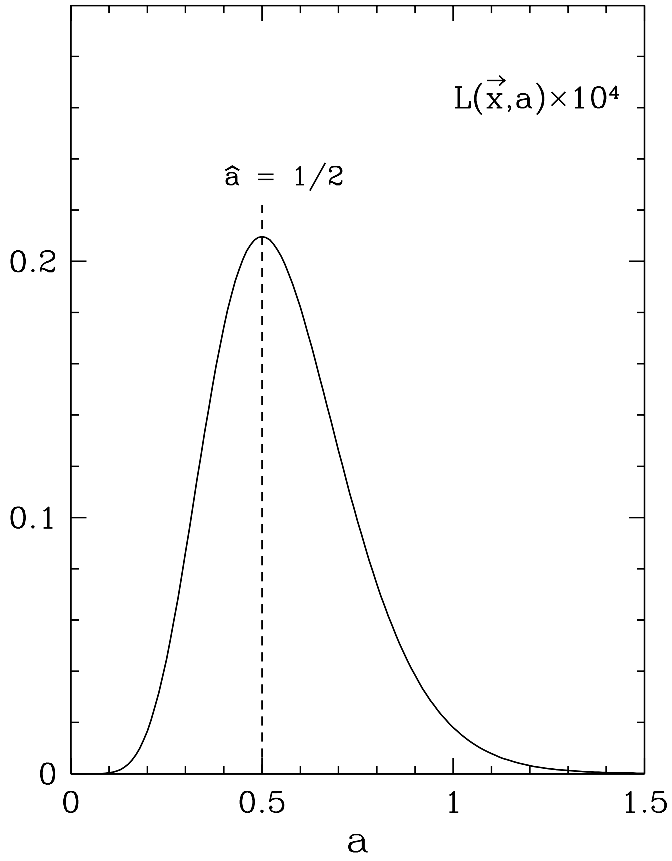
The likelihood function is

$$L(\vec{x}, a) \;=\; \prod_{i=1}^{n} a^2 x_i e^{-a x_i}.$$

The best estimate of a occurs at the maximum of $L(\vec{x}, a)$, which occurs at

$$\left.\frac{\partial L(\vec{x}, a)}{\partial a}\right|_{\hat{a}} = 0 \quad \Rightarrow \left[\begin{matrix}\text{a bit of}\\ \text{algebra}\end{matrix}\right] \Rightarrow \quad \hat{a} = \frac{2n}{\sum x_i}$$

If $n = 4$ and $\vec{x} = (2, 4, 5, 5)$, then $\hat{a} = 1/2$.



The Log-Likelihood Function

For computational convenience, one often prefers to deal with the log of the likelihood function in maximum likelihood calculations.

This is okay because the maxima of the likelihood and its log occur at the same value of the parameters.

The log-likelihood is defined to be

$$\ell(\vec{\mathbf{x}}, \vec{\mathbf{a}}) \;=\; \ln\left\{\mathbf{L}(\vec{\mathbf{x}}, \vec{\mathbf{a}})\right\}$$

and the likelihood equations become

$$\left.\frac{\partial \ell}{\partial \mathbf{a_j}}\right|_{\hat{\mathbf{a}}} \;=\; \mathbf{0}.$$

A Fully Realistic Example - 1

We have n independent measurements (x_i, σ_i) drawn from the Gaussians

$$f_i(\xi_i, \sigma_i, a) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_i}\,\exp\left[-\frac{1}{2}\,\frac{(\xi_i - a)^2}{\sigma_i^2}\right].$$

Thus, the measurements all have the same mean value but have different noise. The noise is described by the width of the Gaussians, a different width for each measurement.

The joint probability distribution for the data points is

$$f(\vec{\xi}, \vec{\sigma}, a) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i}\,\exp\left[-\frac{1}{2}\,\frac{(\xi_i - a)^2}{\sigma_i^2}\right],$$

The joint likelihood function for all the measurements is

$$L(\vec{x}, \vec{\sigma}, a) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i}\,\exp\left[-\frac{1}{2}\,\frac{(x_i - a)^2}{\sigma_i^2}\right].$$

A Fully Realistic Example - 2

The log-likelihood function is

$$\ell(\vec{\mathbf{x}}, \vec{\sigma}, \mathbf{a}) \;=\; \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma_i}\right) \;-\; \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - a)^2}{\sigma_i^2}.$$

From elementary probability theory we know that the mean value of the Gaussian is a. Let us estimate the mean value.

The likelihood equation for â is

$$0 \;=\; \left.\frac{\partial \ell}{\partial a}\right|_{\hat{a}} \;=\; \frac{\partial}{\partial a}\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - a)^2}{\sigma_i^2}\right]_{\hat{a}}.$$

After some algebra we arrive at

$$\hat{a} \;=\; \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \qquad \text{where } w_i = 1/\sigma_i^2$$

This is the same as the weighted average in freqentist statistics!

Fit a Model to Data - 1

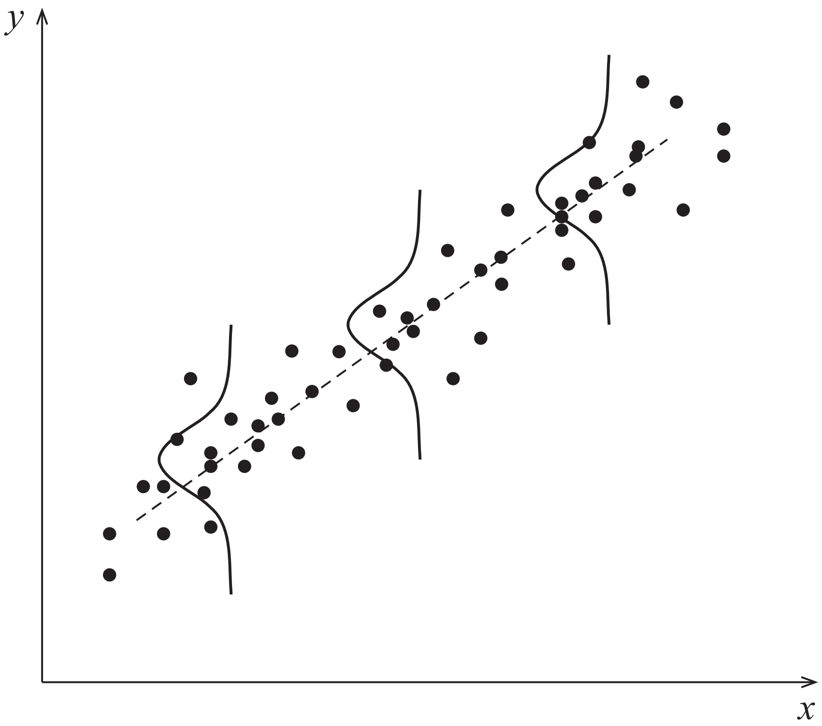
Have independent data points:

$$(\mathbf{x_i}, \mathbf{y_i}, \sigma_\mathbf{i}), \qquad \mathbf{i = 1, \cdots, n}$$

The $\mathbf{y_i}$ have errors $\sigma_\mathbf{i}$.

We know the $\mathbf{y_i}$ are drawn from Gaussian distributions

$$\mathbf{f(y)} \; \propto \; \exp\left[-\frac{(\mathbf{y}-\mu)^2}{\mathbf{2}\sigma^2}\right].$$



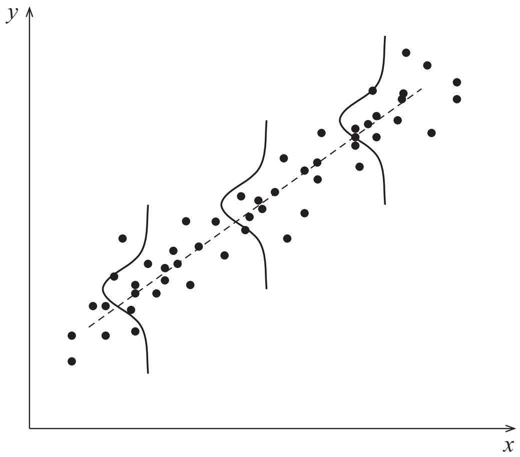
To represent the data both μ and σ must depend on x.

- The values of σ are given explicitly for each $\mathbf{x_i}$.
- The values of μ are given as a function of x:

$$\mu = \mu(\mathbf{x})$$

Fit a Model to Data - 2

The functional form of $\mu(\mathrm{x})$ can be as complicated as desired.



To keep the example simple, fit a straight line to the data:

$$\mu = \mathbf{a_0 + a_1 x}$$

The individual $\mathrm{y_i}$ are now each drawn from Gaussian distributions

$$\mathbf{f(y)} \propto \exp\left[-\frac{(\mathbf{y - a_0 - a_1 x_i})^2}{2\sigma_i^2}\right].$$

Because the data points are independent of each other, the joint probability distribution is the product of the individual distributions. The likelihood function is, therefore,

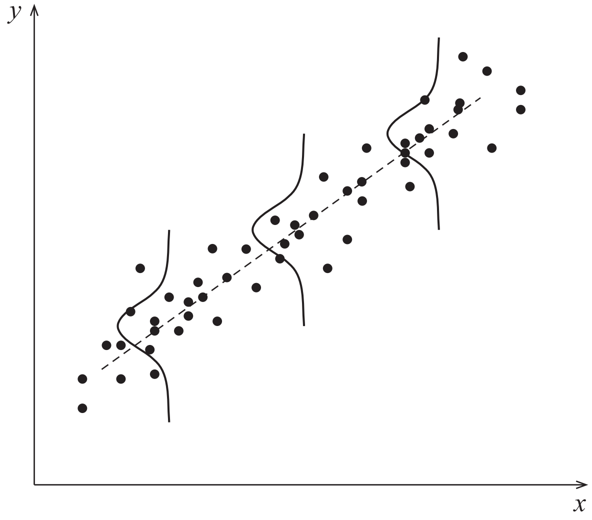
$$\mathbf{L}(\vec{\mathbf{x}}, \vec{\mathbf{y}}, \vec{\sigma}, \mathbf{a_0}, \mathbf{a_1}) \propto \prod_{\mathrm{i=1}}^{\mathrm{n}} \exp\left[-\frac{(\mathbf{y_i - a_0 - a_1 x_i})^2}{2\sigma_i^2}\right].$$

Fit a Model to Data - 3

The log likelihood function is

$$\ell(\vec{\mathbf{x}}, \vec{\mathbf{y}}, \vec{\sigma}, \mathbf{a}_0, \mathbf{a}_1) \;=\; -\frac{1}{2}\sum_{i=1}^{n} \mathbf{w}_i(\mathbf{y}_i - \mathbf{a}_0 - \mathbf{a}_1\mathbf{x}_i)^2 + \mathbf{c},$$

where $\mathbf{w}_i = 1/\sigma_i^2$ and c is a constant.



The two likelihood equations are

$$\left.\frac{\partial\ell}{\partial \mathbf{a}_0}\right|_{\vec{\mathbf{a}}} = 0 \quad \text{and} \quad \left.\frac{\partial\ell}{\partial \mathbf{a}_1}\right|_{\vec{\mathbf{a}}} = 0,$$

which lead to the following equations (in matrix form) for $\hat{\mathbf{a}}_0$ and $\hat{\mathbf{a}}_1$:

$$\begin{pmatrix} \sum_i \mathbf{w}_i & \sum_i \mathbf{w}_i\mathbf{x}_i \\ \sum_i \mathbf{w}_i\mathbf{x}_i & \sum_i \mathbf{w}_i\mathbf{x}_i^2 \end{pmatrix}\begin{pmatrix} \hat{\mathbf{a}}_0 \\ \hat{\mathbf{a}}_1 \end{pmatrix} = \begin{pmatrix} \sum_i \mathbf{w}_i\mathbf{y}_i \\ \sum_i \mathbf{w}_i\mathbf{x}_i\mathbf{y}_i \end{pmatrix}.$$

These are identical to the normal equations for a weighted least squares (or χ^2 minimization) solution for $\hat{\mathbf{a}}_0$ and $\hat{\mathbf{a}}_1$.

Comparison to Standard Frequentist Statistics

- The probability distributions from which the data points are drawn *must* be known to apply likelihood statistics, but not for many standard frequentist techniques.
- If the data have Gaussian distributions, likelihood statistics reduces to ordinary frequentist statistics.
- Likelihood statistics provides a solid foundation for treating data with non-Gaussian distributions (e.g., the Poisson distribution in some astronomical applications).
- If treated as probability distributions, likelihood functions can be analyzed with all the tools developed to analyze posterior distributions of Bayesian statistics (e.g., marginal distributions and MCMC sampling).

Comparison to Bayesian Statistics

The salient feature of Bayesian statistics: Combine new data with existing knowledge using Bayes' equation:

$$\text{(Posterior Probabiltiy)} \;\propto\; \text{(Likelihood)} \times \text{(Prior Probability)}.$$

- Mathematically, likelihood statistics is essentially Bayesian statistics without a prior probability distribution.

$$\begin{aligned} \text{Likelihood function} &\iff \text{Posterior distribution} \\ \text{Likelihood ratio} &\iff \text{Bayes factor} \end{aligned}$$

- It is *not* Bayesian statistics with a flat or uninformative prior.
  - Flatness is not an invariant concept.
  - The prior must "know" about the likelihood function to be truly uninformative.
- Likelihood statistics defines probability as a frequency, not as a Bayesian state of knowledge or state of belief.