3

Statistics Background

We review some standard concepts in mathematical statistics, such as likelihood, sufficiency and Bayesian statistics. We describe a variety of Monte Carlo methods, and work out how to deal with partially observed models in both frequentist and Bayesian settings.

3.1 Maximum Likelihood Theory

In this section, we review results from the maximum likelihood theory. Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be a random vector of observations and let $\mathscr{L}(\mathbf{y}, \boldsymbol{\theta})$ be the likelihood of observing $\mathbf{y}$ given parameter vector $\boldsymbol{\theta}$. Often, it is understood that the likelihood must be a function of data and $\mathbf{y}$ is dropped from the notation: $\mathscr{L}(\mathbf{y}, \boldsymbol{\theta}) = \mathscr{L}(\boldsymbol{\theta})$. The method of maximum likelihood prescribes to estimate $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathscr{L}(\boldsymbol{\theta}),$$

where $\boldsymbol{\Theta}$ is the parameter space. Maximizing $\mathscr{L}(\boldsymbol{\theta})$ is equivalent to maximizing the log likelihood $l(\boldsymbol{\theta}) = \ln \mathscr{L}(\boldsymbol{\theta})$. The latter maximization can be more convenient when done analytically and numerically more stable when done numerically, because $\mathscr{L}(\boldsymbol{\theta})$ is usually a product with a large number of terms.

Example: *Multinomial sampling* Let $y_1, \ldots, y_n$ be iid random variables with $\Pr(y_i = j) = p_j$ for $j = 1, \ldots, s$ and $i = 1, \ldots, n$ such that $\sum_{j=1}^{s} p_j = 1$. Suppose that we observe $n_j = \sum_{i=1}^{n} 1_{\{y_i = j\}}$ for $j = 1, \ldots, s$. Then $(n_1, \ldots, n_s)$ is multinomially distributed with parameters $\mathbf{p} = (p_1, \ldots, p_s)$.

Likelihood:

$\mathscr{L}(\mathbf{p}) = \frac{n!}{n_1! \cdots n_s!} p_1^{n_1} \cdots p_s^{n_s}$, where $n = \sum_{j=1}^{s} n_j$.

log likelihood:

$l(\mathbf{p}) = \ln\left(\frac{n!}{n_1! \cdots n_s!}\right) + \sum_{j=1}^{s} n_j \ln p_j$.

Objective:

maximize $l(\mathbf{p}) = \sum_{j=1}^{s} n_j \ln p_j$ subject to the constraint imposed by the parameter space $\boldsymbol{\Theta} = \{p_j \geq 0; \sum_j p_j = 1\}$. We solve this optimization problem by Lagrange multipliers and form an auxiliary objective function

$$F(\mathbf{p}, \lambda) = \sum_{j=1}^{s} n_j \ln p_j + \lambda \left(1 - \sum_{j=1}^{s} p_j\right).$$

Differentiating this function with respect to $p_1, \ldots, p_s$ and λ gives

$$\begin{aligned}
\frac{\partial F}{\partial p_1} &= n_1 \frac{1}{p_1} - \lambda = 0 \Rightarrow n_1 = \lambda p_1 \\
&\vdots \\
\frac{\partial F}{\partial p_s} &= n_s \frac{1}{p_s} - \lambda = 0 \Rightarrow n_s = \lambda p_s \\
\frac{\partial F}{\partial \lambda} &= 1 - \sum_{j=1}^{s} p_j
\end{aligned}$$

Summing the first s equations we get $n = \sum_{j=1}^{s} n_j = \lambda \sum_{j=1}^{s} p_j = \lambda$. Therefore, $\hat{p}_j = n_j/\lambda = n_j/n$. Checking that this solution is a maximum is left to the reader.

3.1.1 Reminder: types of convergences

1. **Convergence in distribution**: given a sequence of random variables $X_1, X_2, \ldots$, and their corresponding cdfs $F_1(x), F_2(x), \ldots$, we say that X_n converges to X, with cdf $F(x)$, in distribution if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x such that F is continuous at x.
   Notation: $X_n \xrightarrow{\text{D}} X$.
2. **Convergence in probability**: X_n converges to X in probability if $\lim_{n\to\infty} \Pr(|X_n - X| \geq \varepsilon) = 0$ for all $\varepsilon > 0$.
   Notation: $X_n \xrightarrow{\text{P}} X$.
3. **Almost sure convergence**: X_n converges to X almost surely if $\Pr(\lim_{n\to\infty} X_n = X) = 1$.
   Notation: $X_n \xrightarrow{\text{a.s.}} X$.

   3. $\Rightarrow$ 2. $\Rightarrow$ 1.

3.1.2 Properties of ml estimators

Under very mild conditions for iid observations,

1. the mle is consistent, meaning that $\hat{\boldsymbol{\theta}} \xrightarrow{\text{P}} \boldsymbol{\theta}$ as sample size increases to ∞.
2. the mle is asymptotically unbiased, meaning that $\text{E}\left(\hat{\boldsymbol{\theta}}\right) \to \boldsymbol{\theta}$ as sample size increases to ∞, where $\boldsymbol{\theta}$ is the true value of the parameter vector.
3. the mle is asymptotically normally distributed: $\mathbf{I}_n^{1/2}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{D}} \mathcal{N}(\mathbf{0}, \mathbf{I})$, where

   $$\mathbf{I}_n(\boldsymbol{\theta}) = -\text{E}\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \text{ is called } \textbf{Fisher information matrix,}$$

   $\mathbf{0}$ is a vector of zeros and $\mathbf{I}$ is an identity matrix.
   For iid data, $l(\boldsymbol{\theta}) = \sum_i l_i(\boldsymbol{\theta})$, making it possible to define

   $$\mathbf{F}(\boldsymbol{\theta}) = -\text{E}\left(\frac{\partial^2 l_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right).$$

   Then asymptotic normality can be expressed as $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{D}} \mathcal{N}(\mathbf{0}, \mathbf{F}^{-1})$,

Example: *Trinomial sampling* Let us first get rid of the constraint to make our derivations cleaner so that the log likelihood becomes

$$l(\mathbf{p}) = n_1 \ln(1-p_2-p_3) + n_2 \ln p_2 + n_3 \ln p_3 + C$$

We have shown that $\hat{p}_j = n_j/n$. $\mathrm{E}(\hat{p}_j) = np_j/n = p_j$, therefore our estimator is unbiased even for finite samples. This is not always true for an arbitrary mle.

The first order partial derivatives of the log likelihood are

$$\frac{\partial l}{\partial p_2} = -\frac{n_1}{1-p_2-p_3} + \frac{n_2}{p_2}, \qquad \frac{\partial l}{\partial p_3} = -\frac{n_1}{1-p_2-p_3} + \frac{n_3}{p_3}.$$

The second order partial derivatives are

$$\begin{aligned}
\frac{\partial^2 l}{\partial p_2^2} &= -\frac{n_1}{(1-p_2-p_3)^2} - \frac{n_2}{p_2^2}, \\
\frac{\partial^2 l}{\partial p_3^2} &= -\frac{n_1}{(1-p_3-p_3)^2} - \frac{n_3}{p_3^2}, \\
\frac{\partial^2 l}{\partial p_2 \partial p_3} &= -\frac{n_1}{(1-p_2-p_3)^2}.
\end{aligned}$$

The Fisher information matrix becomes

$$\mathbf{I}_n(p_2,p_3) = \mathrm{E}\begin{pmatrix} \frac{n_1}{(1-p_2-p_3)^2} + \frac{n_2}{p_2^2} & \frac{n_1}{(1-p_2-p_3)^2} \\ \frac{n_1}{(1-p_2-p_3)^2} & \frac{n_1}{(1-p_2-p_3)^2} + \frac{n_3}{p_3^2} \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{E}(n_1)}{(1-p_2-p_3)^2} + \frac{\mathrm{E}(n_2)}{p_2^2} & \frac{\mathrm{E}(n_1)}{(1-p_2-p_3)^2} \\ \frac{\mathrm{E}(n_1)}{(1-p_2-p_3)^2} & \frac{\mathrm{E}(n_1)}{(1-p_2-p_3)^2} + \frac{\mathrm{E}(n_3)}{p_3^2} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{n(1-p_2-p_3)}{(1-p_2-p_3)^2} + \frac{np_2}{p_2^2} & \frac{n(1-p_2-p_3)}{(1-p_2-p_3)^2} \\ \frac{n(1-p_2-p_3)}{(1-p_2-p_3)^2} & \frac{n(1-p_2-p_3)}{(1-p_2-p_3)^2} + \frac{n_3}{p_3^2} \end{pmatrix} = n\begin{pmatrix} \frac{1-p_3}{p_1 p_2} & \frac{1}{p_1} \\ \frac{1}{p_1} & \frac{1-p_2}{p_1 p_3} \end{pmatrix}$$

The inverse of the Fisher information is

$$\mathbf{I}_n^{-1}(p_2,p_3) = \frac{1}{n}\begin{pmatrix} p_2(1-p_2) & -p_2 p_3 \\ -p_2 p_3 & p_3(1-p_3). \end{pmatrix}$$

Now consider the covariance matrix of $(\hat{p}_2, \hat{p}_3)$:

$$\mathrm{Var}\begin{pmatrix} \hat{p}_2 \\ \hat{p}_3 \end{pmatrix} = \mathrm{Var}\begin{pmatrix} n_2/n \\ n_3/n \end{pmatrix} = \mathbf{I}_n^{-1}(p_2,p_3).$$

Again, in this particular case the result is valid even for finite samples. In general, we use asymptotics to *approximate* the covariance of mle with the inverse Fisher information matrix.

Note 3.1. Since $\boldsymbol{\theta}$ is never known to us, in practice we plug in $\hat{\boldsymbol{\theta}}$ into the Fisher information matrix.

3.1.3 Likelihood ratio test

Suppose we we chose to work with a parametric log likelihood $l(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If we want to test the hypothesis $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$, we first obtain mles under restricted and unrestricted models:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_0 &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} l(\boldsymbol{\theta}), \\
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l(\boldsymbol{\theta}).
\end{aligned}$$

Notice that $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$ implies that $l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}) \geq 0$. Moreover, if H_0 is true, then asymptotically $2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0] \sim \chi_q^2$, where $q = \dim(\boldsymbol{\Theta}) - \dim(\boldsymbol{\Theta}_0)$. We can use this distribution to decide when the difference between the two log likelihoods is too large.

3.2 Sufficiency

A statistic $T(X_1,\ldots,X_n)$ is called sufficient for $\boldsymbol{\theta}$ if the conditional density (or probability function) of the data, given the value of $T(X_1,\ldots,X_n)$, does not depend on $\boldsymbol{\theta}$. In other words, if you tell us the value of $T(X_1,\ldots,X_n)$, we don't need to know anything else about the sample in order to draw inference about the value of θ.

Example: *Binomial case* Let X_1, X_2, X_3 be iid Bin$(1,\theta)$, and let $T(\mathbf{X}) = (X_1+2X_2+3X_3)/6$. Do we need to know more about the sample than just the value of T in order to make a good guess as to the value of θ? To check that, suppose that the sample is $(X_1,X_2,X_3) = (1,1,0)$, so that the observed value of T is $1/2$. Then

$$\Pr(\mathbf{X}=(1,1,0)\,|\,T(\mathbf{X})=1/2) = \frac{\Pr(\mathbf{X}=(1,1,0)\cap T(\mathbf{X})=1/2)}{P(T(\mathbf{X})=1/2)}$$
$$= \frac{\Pr(\mathbf{X}=(1,1,0)}{P(T(\mathbf{X})=1/2)} = \frac{\theta^2(1-\theta)}{\Pr(\mathbf{X}=(1,1,0)\cup T(\mathbf{X})=1/2)} = \frac{\theta^2(1-\theta)}{\theta^2(1-\theta)+\theta(1-\theta)^2} = \theta.$$

Since this probability depends on θ, we need more information about the sample than just the fact that $T(\mathbf{X}) = 1/2$. When $T = 1/2$ there are two possible explanations, either $X = (1,1,0)$, which would be likely if θ were large, or $X = (0,0,1)$, which would be likely if θ were small. Thus, the information that $T = 1/2$ does not tell us much about the actual value of θ.

The method used in this example is fine when we want to show that a statistic is not sufficient, but it is not all that helpful in trying to figure out reasonable candidates for sufficient statistics. A criterion for doing this is the following, due to Fisher and Neyman.

Proposition 3.1. *(Fisher-Neyman factorization criterion)*
A statistic $T(\mathbf{X})$ is sufficient if and only if the density (or probability function) can be factored as

$$f(\mathbf{x};\boldsymbol{\theta}) = g(T(\mathbf{x});\boldsymbol{\theta})h(\mathbf{x}),$$

where g only depends on the x_i through $T(\mathbf{x})$, while h does not depend on θ.

Proposition 3.2. *If $T(\mathbf{X})$ is a sufficient statistic under a particular probability model, then the maximum likelihood estimators of the model parameters and the observed Fisher information matrix depend on the data only through $T(\mathbf{X})$.*

The above proposition implies that for the purposes of finding mle's and forming confidence intervals via the observed information matrix, we don't lose any information available in the observed data by compressing them into sufficient statistics.

Example: *Binomial case, continued* The probability mass function of $\mathbf{X}$ is

$$P(\mathbf{X}=\mathbf{x}) = \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\theta^{x_3}(1-\theta)^{1-x_3} = \left(\frac{\theta}{1-\theta}\right)^{\Sigma_{i=1}^3 x_i}(1-\theta)^3.$$

Using for g the entire expression on the right-hand side, and letting $h(x) = 1$, we see that g depends on the data $\mathbf{x}$ only through their sum $\sum x_i$, which therefore is a sufficient statistic. To check back with the definition, notice that $\sum_{i=1}^3 X_i \sim \text{Bin}(3,\theta)$, so that

$$\Pr\left(\mathbf{X}=\mathbf{x}\,|\,\sum_{i=1}^3 X_i = \sum_{i=1}^3 x_i\right) = \frac{\theta^{\Sigma_{i=1}^3 x_i}(1-\theta)^{n-\Sigma_{i=1}^3 x_i}}{\binom{3}{\Sigma_{i=1}^3 x_i}\theta^{\Sigma_{i=1}^3 x_i}(1-\theta)^{n-\Sigma_{i=1}^3 x_i}} = \frac{1}{\binom{3}{\Sigma_{i=1}^3 x_i}}.$$

Since the right-hand side is independent of θ, regardless of the values of the x_i, we see that $\sum X_i$ is indeed a sufficient statistic. Notice that, in fact, the conditional distribution is uniform over the set of possible outcomes with the given value of the sufficient statistic.

Example: *The normal case* Suppose now that $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. First assume that σ^2 is a known number, so that we are only interested in estimating μ. Then

$$f(\mathbf{x};\mu) = \left(\sigma\sqrt{2\pi}\right)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right) = \exp\left(\frac{\mu\sum x_i}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)\exp\left(-\frac{\sum x_i^2}{2\sigma^2} + n\log\left(\sigma\sqrt{2\pi}\right)\right).$$

The first exponential function is g, the second is h. We see that $\sum X_i$ is a sufficient statistic.

Now assume instead that μ is known, and σ is the parameter of interest. Then we write the density

$$f(\mathbf{x};\sigma) = \left(\sigma\sqrt{2\pi}\right)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right).$$

In this case $\sum(X_i - \mu)^2$ is a sufficient statistic. We have $h(x) = 1$.

Finally, if both parameters are unknown, we write the density

$$f(\mathbf{x};\mu,\sigma) = \exp\left(\frac{\mu\sum x_i}{\sigma^2} - \frac{\sum x_i^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + n\log\left(\sigma\sqrt{2\pi}\right)\right),$$

from which it follows that $\left(\sum X_i, \sum X_i^2\right)$ together are sufficient for (μ, σ^2).

3.3 Bayesian Inference

We start with observed data $\mathbf{y}$ and their sampling density $\Pr(\mathbf{y} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are model parameters. We will abuse notation by using $\Pr(\cdot)$ to denote sampling densities, prior densities, and other related objects. We also need a prior distribution for $\boldsymbol{\theta}$, $\Pr(\boldsymbol{\theta})$. The posterior distribution of $\boldsymbol{\theta}$ is derived via Bayes' rule:

$$\Pr(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\Pr(\mathbf{y}, \boldsymbol{\theta})}{\Pr(\mathbf{y})} = \frac{\Pr(\mathbf{y} \mid \boldsymbol{\theta})\Pr(\boldsymbol{\theta})}{\int \Pr(\mathbf{y} \mid \boldsymbol{\theta})\Pr(\boldsymbol{\theta})\, d\boldsymbol{\theta}},$$

where $\Pr(\mathbf{y})$ is called a marginal or integrated likelihood. Bayesian estimators are formed by summarizing the posterior distribution in a meaningful way, formally, using Bayesian decision theory. For example, one Bayesian estimator is the posterior mean:

$$\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta} \mid \mathbf{y}) = \int \boldsymbol{\theta}\Pr(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}.$$

In practice, posterior mean, median, and mode are the most commonly used Bayesian estimators.

Note 3.2. Bayesian and frequentist statistical approaches differ not in how these two schools of thought arrive at estimators, but rather in their approaches of evaluating uncertainty of estimators and hypothesis testing, as a result. A Bayesian statistician will use only the posterior distribution to give statements about how likely the true parameter value is to be in a certain region of the parameter space. However, a frequentist statistician should be perfectly happy to work with frequentist properties of any Bayesian estimator.

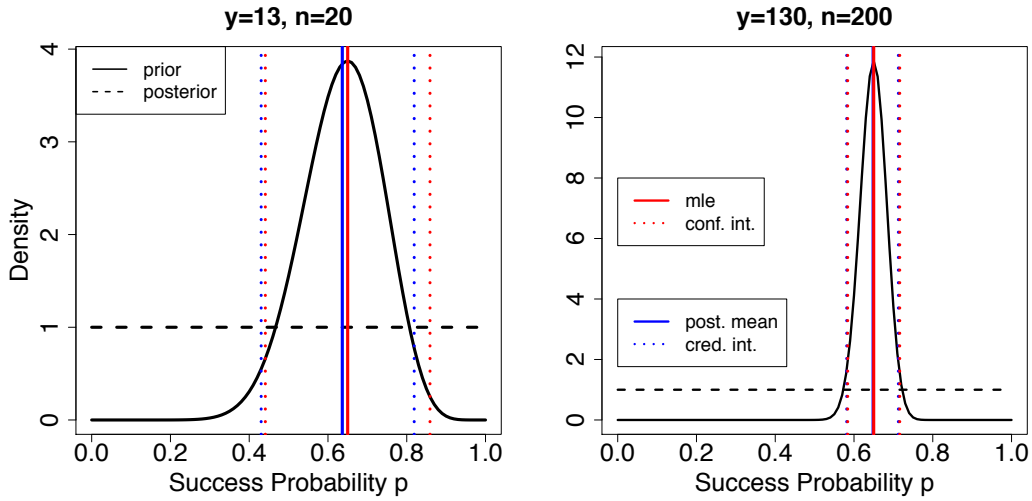

FIGURE 3.1: Maximum likelihood and Bayesian analysis of a binomial experiment.

Example: *Binomial likelihood with beta prior* Let $y \sim \text{Bin}(n,p)$ so that the likelihood is

$$\Pr(y\,|\,p) = \binom{n}{y} p^y (1-p)^{n-y}.$$

We assume a beta prior distribution for the success probability p:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

The posterior distribution is

$$\Pr(p\,|\,y) = C p^{x+\alpha-1}(1-p)^{n-x+\beta-1},$$

where normalization constant can be worked out as

$$\int \Pr(p\,|\,y)\,dp = 1 \Rightarrow C\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} = 1.$$

Therefore,

$$\Pr(p\,|\,y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

which means that $p\,|\,y \sim \text{Beta}(y+\alpha, n-y+\beta)$. The posterior mean becomes:

$$\text{E}(p\,|\,y) = \frac{y+\alpha}{n-y+\beta+x+\alpha} = \frac{y+\alpha}{n+\alpha+\beta}.$$

Compare this Bayesian estimator to the mle: y/n. If we want to form Bayesian credible intervals, we simply need to compute 2.5% and 97.5% quantiles of the posterior Beta density. Recall, that asymptotic confidence interval for mle is $(\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n})$. We plot all these quantities for $\alpha = \beta = 1$, corresponding to the uniform distribution for p, and for $y = 13, n = 20$, and $y = 130, 200$ in Figure 3.1.

Note 3.3. When we start with a parametric family of prior distributions and the corresponding posterior distribution remains in this family, we say that such a prior is **conjugate** to the likelihood. We have seen that beta prior is conjugate to the binomial likelihood. Similarly Dirichlet prior is conjugate to the multinomial likelihood. More specifically, if

$$(p_1,\ldots,p_s) \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_s) \text{ and}$$
$$(n_1,\ldots,n_s) \sim \text{Multinomial}(p_1,\ldots,p_s),$$

then

$$p_1,\ldots,p_s \mid n_1,\ldots,n_2 \sim \text{Dirichlet}(\alpha_1+n_1,\ldots,\alpha_s+n_s).$$

Other frequently used examples of conjugate prior-likelihood pairs include: normal-normal and gamma-Poisson families.

Proposition 3.3. *If $T(\mathbf{X})$ is a sufficient statistic under a particular probability model, then regardless of the prior specification, the posterior distribution of the model parameters depends on the data only through $T(\mathbf{X})$.*

The above proposition implies that (also in the Bayesian case) there is no loss of information available in the observed data by compressing them into sufficient statistics.

3.3.1 Bayes factors

In Bayesian framework, two arbitrary models M_1 and M_2 can be compared using their integrated likelihoods. Suppose $\Pr(\mathbf{y} \mid \boldsymbol{\theta}_{M_1}, M_1)$ and $\Pr(\mathbf{y} \mid \boldsymbol{\theta}_{M_2}, M_2)$ are likelihoods of the two models and $\Pr(\boldsymbol{\theta}_{M_1})$ and $\Pr(\boldsymbol{\theta}_{M_2})$ are corresponding priors. Then Bayes factor is computed as

$$\text{BF}_{12} = \frac{\Pr(\mathbf{y} \mid M_1)}{\Pr(\mathbf{y} \mid M_2)} = \frac{\int \Pr(\mathbf{y} \mid \boldsymbol{\theta}_{M_1}, M_1)\Pr(\boldsymbol{\theta}_{M_1}) d\boldsymbol{\theta}_{M_1}}{\int \Pr(\mathbf{y} \mid \boldsymbol{\theta}_{M_2}, M_2)\Pr(\boldsymbol{\theta}_{M_2}) d\boldsymbol{\theta}_{M_2}}.$$

If we assign prior probabilities to M_1 and M_2: $\Pr(M_1)$ and $\Pr(M_2)$, then we can rewrite the Bayes factor as

$$\text{BF}_{12} = \frac{\Pr(\mathbf{y} \mid M_1)}{\Pr(\mathbf{y} \mid M_2)} = \frac{\Pr(\mathbf{y}, M_1)}{\Pr(\mathbf{y}, M_2)} \Big/ \frac{\Pr(M_1)}{\Pr(M_2)} = \underbrace{\frac{\Pr(M_1 \mid \mathbf{y})}{\Pr(M_2 \mid \mathbf{y})}}_{\text{posterior odds}} \Big/ \underbrace{\frac{\Pr(M_1)}{\Pr(M_2)}}_{\text{prior odds}}.$$

See the review in Kass and Raftery (1995) for more details.

3.4 Monte Carlo methods

Although our driving applications of Monte Carlo integration will mostly revolve around Bayesian inference, we would like to point out that all Monte Carlo methods can (should?) be viewed as a numerical integration problem. Such problems usually start with either discrete ($\mathbf{x}$) or continuous ($\boldsymbol{\theta}$) vector of random variables. Despite the fact that distributions of these vectors are known only up to a proportionality constant, we are interested in taking expectations with respect to these distributions. Compare the following integration problems faced by physicists and Bayesian statisticians.

Statistical mechanics	*Bayesian statistics*
$\Pr(\mathbf{x}) = \frac{1}{Z} e^{-\mathscr{E}(\mathbf{x})}$	$\Pr(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{1}{C} \Pr(\mathbf{y} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$
Objective: $\text{E}[f(\mathbf{x})] = \sum_{\mathbf{x}} f(\mathbf{x}) \Pr(\mathbf{x})$	Objective: $\text{E}[f(\boldsymbol{\theta}) \mid \mathbf{y}] = \int f(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta} \mid \mathbf{y}) \, \text{d}\boldsymbol{\theta}$