Principles of Bayesian Inference

Sudipto Banerjee[1] and Andrew O. Finley[2]

[1] Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

[2] Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

March 4, 2013

- We like to think of "Probability" formally as a *function* that assigns a real number to an *event*.
- Let E and F be any events that might occur under an experimental setup H. Then a probability function $P(E)$ is defined as:
  - P1 $0 \leq P(E) \leq 1$ for all E.
  - P2 $P(H) = 1$.
  - P3 $P(E \cup F) = P(E) + P(F)$ whenever it is impossible for any two of the events E and F to occur. Usually consider: $E \cap F = \{\phi\}$ and say they are *mutually exclusive*.

- If E is an event, then we denote its complement ("NOT" E) by $\bar{E}$ or E^c. Since $E \cap \bar{E} = \{\phi\}$, we have from P3:

$$P(\bar{E}) = 1 - P(E).$$

- Conditional Probability of E given F:

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

  Sometimes we write EF for $E \cap F$.

- Compound probability rule: write the above as

$$P(E \mid F)P(F) = P(EF).$$

- Independent Events: E and F are said to be independent if the occurrence of one does not imply the occurrence of the other. Then, $P(E \mid F) = P(E)$ and we have the following *multiplication rule*:

$$P(EF) = P(E)P(F).$$

- If $P(E \mid F) = P(E)$, then $P(F \mid E) = P(F)$.

- Marginalization: We can express $P(E)$ by "marginalizing" over the event F:

$$\begin{aligned} P(E) &= P(EF) + P(E\bar{F}) \\ &= P(F)P(E \mid F) + P(\bar{F})P(E \mid \bar{F}). \end{aligned}$$

Bayes Theorem

- Observe that:

$$P(EF) = P(E \mid F)P(F) = P(F \mid E)P(E)$$
$$\Rightarrow P(F \mid E) = \frac{P(F)P(E \mid F)}{P(E)}.$$

This is ***Bayes' Theorem***, named after Reverend Thomas Bayes – an English clergyman with a passion for gambling!

- Often this is written as:

$$P(F \mid E) = \frac{P(F)P(E \mid F)}{P(F)P(E \mid F) + P(\bar{F})P(E \mid \bar{F})}.$$

- Two hypothesis: H_0 : excess relative risk for thrombosis for women taking a pill exceeds 2; H_1 : it is under 2.
- Data collected at hand from a controlled trial show a relative risk of $x = 3.6$.
- Probability or likelihood under the data, given our prior beliefs is $P(x \mid H)$; H is H_0 or H_1.

- Bayes Theorem *updates* the probability of each hypothesis:

$$P(H \mid x) = \frac{P(H)P(x \mid H)}{P(x)};\ H \in \{H_0, H_1\}\ .$$

- Marginal probability:

$$P(x) = P(H_0)P(x \mid H_0) + P(H_1)P(x \mid H_1)$$

- Reexpress:

$$P(H \mid x) \propto P(H)P(x \mid H);\ H \in \{H_0, H_1\}\ .$$

Likelihood and Prior

- Bayes theorem in English:

$$\text{Posterior distribution} = \frac{\text{prior} \times \text{likelihood}}{(\sum \text{prior} \times \text{likelihood})}$$

- Denominator is summed over *all possible priors*
- It is a fixed normalizing factor that is (usually) extremely difficult to evaluate: curse of dimensionality
- Markov Chain Monte Carlo to the rescue!
- WinBUGS software:
  `www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`

An Example

- Clinician interested in π: proportion of children between age 5–9 in a particular population having asthma symptoms.
- Clinician has prior beliefs about π, summarized as “Prior support” and “Prior weights”
- Data: random sample of 15 children show 2 having asthma symptoms. Likelihood obtained from Binomial distribution:

$$\binom{15}{2}\pi^2(1-\pi)^{13}$$

- Note: $\binom{15}{2}$ is a “constant” and *can* be ignored in the computations (though they are accounted for in the next Table).

Prior Support	Prior weight	Likelihood	Prior $\times$ Likelihood	Posterior
0.10	0.10	0.267	0.027	0.098
0.12	0.15	0.287	0.043	0.157
0.14	0.25	0.290	0.072	0.265
0.16	0.25	0.279	0.070	0.255
0.18	0.15	0.258	0.039	0.141
0.20	0.10	0.231	0.023	0.084
Total	**1.00**		**0.274**	**1.000**

- Posterior: obtained by dividing Prior $\times$ Likelihood with *normalizing constant* **0.274**

- Classical statistics: model parameters are *fixed* and *unknown*.
- A Bayesian thinks of parameters as random, and thus having distributions (just like the data). We can thus think about unknowns for which no reliable frequentist experiment exists, e.g. $\theta =$ proportion of US men with untreated prostate cancer.
- A Bayesian writes down a *prior* guess for parameter(s) $\boldsymbol{\theta}$, say $p(\boldsymbol{\theta})$. He then combines this with the information provided by the observed data $\mathbf{y}$ to obtain the *posterior* distribution of $\boldsymbol{\theta}$, which we denote by $p(\boldsymbol{\theta} \mid \mathbf{y})$.
- All statistical inferences (point and interval estimates, hypothesis tests) then follow from posterior summaries. For example, the posterior means/medians/modes offer point estimates of $\boldsymbol{\theta}$, while the quantiles yield credible intervals.

- The key to Bayesian inference is “learning” or “updating” of prior beliefs. Thus, posterior information $\geq$ prior information.

- Is the classical approach wrong? That may be a controversial statement, but it certainly is fair to say that the classical approach is limited in scope.

- The Bayesian approach expands the class of models and easily handles:
  - repeated measures
  - unbalanced or missing data
  - nonhomogenous variances
  - multivariate data

  – and many other settings that are precluded (or much more complicated) in classical settings.

- We start with a model (likelihood) $f(\mathbf{y} \mid \boldsymbol{\theta})$ for the observed data $\mathbf{y} = (y_1, \ldots, y_n)'$ given unknown parameters $\boldsymbol{\theta}$ (perhaps a collection of several parameters).

- Add a prior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of hyper-parameters.

- The posterior distribution of $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda}) = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times f(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y} \mid \boldsymbol{\lambda})} = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \times f(\mathbf{y} \mid \boldsymbol{\theta})}{\int f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) d\boldsymbol{\theta}}.$$

  We refer to this formula as *Bayes Theorem*.

- Calculations (numerical and algebraic) are usually required only up to a proportionaly constant. We, therefore, write the posterior as:

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\lambda}) \propto p(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \times f(\mathbf{y} \,|\, \boldsymbol{\theta}).$$

- If $\boldsymbol{\lambda}$ are known/fixed, then the above represents the desired posterior. If, however, $\boldsymbol{\lambda}$ are unknown, we assign a prior, $p(\boldsymbol{\lambda})$, and seek:

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \mathbf{y}) \propto p(\boldsymbol{\lambda}) p(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) f(\mathbf{y} \,|\, \boldsymbol{\theta}).$$

  The proportionality constant does not depend upon $\boldsymbol{\theta}$ or $\boldsymbol{\lambda}$:

$$\frac{1}{p(\mathbf{y})} = \frac{1}{\int p(\boldsymbol{\lambda}) p(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) f(\mathbf{y} \,|\, \boldsymbol{\theta}) d\boldsymbol{\lambda} d\boldsymbol{\theta}}$$

- The above represents a *joint* posterior from a *hierarchical model*. The *marginal* posterior distribution for $\boldsymbol{\theta}$ is:

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) = \int p(\boldsymbol{\lambda}) p(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) f(\mathbf{y} \,|\, \boldsymbol{\theta}) d\boldsymbol{\lambda}.$$

- Example: Consider a single data point y from a Normal distribution: $y \sim N(\theta, \sigma^2)$; assume σ is *known*.

$$f(y|\theta) = N(y \,|\, \theta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(y-\theta)^2)$$

- $\theta \sim N(\mu, \tau^2)$, i.e. $p(\theta) = N(\theta \,|\, \mu, \tau^2)$; μ, τ^2 are known.
- Posterior distribution of θ

$$\begin{aligned} p(\theta|y) &\propto N(\theta \,|\, \mu, \tau^2) \times N(y \,|\, \theta, \sigma^2) \\ &= N\left(\theta \,|\, \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\mu + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}y, \frac{1}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}\right) \\ &= N\left(\theta \,|\, \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right). \end{aligned}$$

- Interpret: Posterior mean is a weighted mean of prior mean and data point.
- The direct estimate is shrunk towards the prior.
- What if you had n observations instead of one in the earlier set up? Say $\mathbf{y} = (y_1, \ldots, y_n)'$, where $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$.
- $\bar{y}$ is a *sufficient statistic* for θ; $\bar{y} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$
- Posterior distribution of θ

$$
\begin{aligned}
p(\theta \,|\, \mathbf{y}) &\propto N(\theta \,|\, \mu, \tau^2) \times N\left(\bar{y} \,|\, \theta, \frac{\sigma^2}{n}\right) \\
&= N\left(\theta \,|\, \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\mu + \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\bar{y}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right) \\
&= N\left(\theta \,|\, \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{y}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)
\end{aligned}
$$

- Consider the problem of estimating the current weight of a group of people. A sample of 10 people were taken and their average weight was calculated as $\bar{y} = 176$ lbs. Assume that the population standard deviation was known as $\sigma = 3$. Assuming that the data $y_1, ..., y_{10}$ came from a $N(\theta, \sigma^2)$ population perform the following:

- Obtain a 95% confidence interval for θ using classical methods.

- Assume a prior distribution for θ of the form $N(\mu, \tau^2)$. Obtain 95% posterior credible intervals for θ for each of the cases: (a) $\mu = 176, \tau = 8$; (b) $\mu = 176, \tau = 1000$ (c) $\mu = 0, \tau = 1000$. Which case gives results closest to that obtained in the classical method?

- Example: Let Y be the number of successes in n independent trials.

$$P(Y = y|\theta) = f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- Prior: $p(\theta) = Beta(\theta|a, b)$:

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

- Prior mean: $\mu = a/(a+b)$; Variance $ab/((a+b)^2(a+b+1))$
- Posterior distribution of θ

$$p(\theta|y) = Beta(\theta|a+y, b+n-y)$$

- Point estimation is easy: simply choose an appropriate distribution summary: posterior mean, median or mode.
- Mode sometimes easy to compute (no integration, simply optimization), but often misrepresents the "middle" of the distribution – especially for one-tailed distributions.
- Mean: easy to compute. It has the "opposite effect" of the mode – chases tails.
- Median: probably the best compromise in being robust to tail behaviour although it may be awkward to compute as it needs to solve:

$$\int_{-\infty}^{\boldsymbol{\theta}_{median}} p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} = \frac{1}{2}.$$

- The most popular method of inference in practical Bayesian modelling is interval estimation using *credible sets*. A $100(1-\alpha)\%$ credible set C for $\boldsymbol{\theta}$ is a set that satisfies:

$$P(\boldsymbol{\theta} \in C \,|\, \mathbf{y}) = \int_C p(\boldsymbol{\theta} \,|\, \mathbf{y}) d\boldsymbol{\theta} \geq 1 - \alpha.$$

- The most popular credible set is the simple equal-tail interval estimate (q_L, q_U) such that:

$$\int_{-\infty}^{q_L} p(\theta \,|\, \mathbf{y}) d\theta = \frac{\alpha}{2} = \int_{q_U}^{\infty} p(\theta \,|\, \mathbf{y}) d\theta$$

  Then clearly $P(\theta \in (q_L, q_U) \,|\, \mathbf{y}) = 1 - \alpha$.

- This interval is relatively easy to compute and has a direct interpretation: The probability that θ lies between (q_L, q_U) is $1 - \alpha$. The frequentist interpretation is extremely convoluted.

- Previous example: direct evaluation of the posterior probabilities. Feasible only for simpler problems.
- Modern Bayesian Analysis: Derive complete posterior densities, say $p(\theta \mid \mathbf{y})$ by drawing samples from that density. Samples are of the parameters themselves, or of their functions.
- If $\theta_1, \ldots, \theta_M$ are samples from $p(\theta \mid \mathbf{y})$ then, densities are created by feeding them into a density plotter. Similarly samples from $f(\theta)$, for some function f, are obtained by simply feeding the θ_i's to $f(\cdot)$.
- In principle M can be arbitrarily large – it comes from the computer and only depends upon the *time* we have for analysis. Do not confuse this with the data sample size n which is limited in size by experimental constraints.