Basic Elements of Bayesian Analysis

In a frequentist analysis, one chooses a model (likelihood function) for the available data, and then either calculates a $p$-value (which tells you how unusual your data would be, assuming your null hypothesis is exactly true), or calculates a confidence interval. We have already seen the many deficiencies of $p$-values, and confidence intervals, while more useful, have a somewhat unnatural interpretation, and ignore any prior information that may be available.

Alternatively, Bayesian inferences can be calculated, which provide direct probability statements about parameters of interest, at the expense of having to first summarize prior information about the parameters of interest.

In a nutshell, the choice between frequentist and Bayesian inferences can be seen as a choice between being “backwards” (frequentists calculate $P(\text{data}|H_0)$, rather than $P(H_0|\text{data})$) or being “subjective” (Bayesian analyses require prior information, which must be elicited subjectively).
The basic elements in a “full” Bayesian analysis are:

1. The parameter of interest, say $\theta$. Note that this is completely general, since $\theta$ may be vector valued. So $\theta$ might be a binomial parameter, or the mean and variance from a Normal distribution, or an odds ratio, or a set of regression coefficients, etc. The parameter of interest is sometimes usefully thought of as the “true state of nature”.

2. The prior distribution of $\theta$, $f(\theta)$. This prior distribution summarizes what is known about $\theta$ before the experiment is carried out. It is “subjective”, so may vary from investigator to investigator.

3. The likelihood function, $f(x|\theta)$. The likelihood function provides the distribution of the data, $x$, given the parameter value $\theta$. So it may be the binomial likelihood, a normal likelihood, a likelihood from a regression equation with associated normal residual variance, logistic regression model, etc.

4. The posterior distribution, $f(\theta|x)$. The posterior distribution summarizes the information in the data, $x$, together with the information in the prior distribution, $f(\theta)$. Thus, it summarizes what is known about the parameter of interest $\theta$ after the data are collected.

5. Bayes Theorem. This theorem relates the above quantities:

$$\text{posterior distribution} = \frac{\text{likelihood of the data} \times \text{prior distribution}}{\text{a normalizing constant}},$$

or

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{\int f(x|\theta) \times f(\theta)d\theta},$$

or, forgetting about the normalizing constant,

$$f(\theta|x) \propto f(x|\theta) \times f(\theta).$$

Thus we “update” the prior distribution to a posterior distribution after seeing the data via Bayes Theorem.
6. The action, \( a \). The action is the decision or action that is taken after the analysis is completed. For example, one may decide to treat a patient with Drug 1 or Drug 2, depending on the data collected in a clinical trial. Thus our action will either be to use Drug 1 (so that \( a = 1 \)) or Drug 2 (so that \( a = 2 \)).

7. The loss function, \( L(\theta, a) \). Each time we choose an action, there is some loss we incur, which depends on what the true state of nature is, and what action we decide to take. For example, if the true state of nature is that Drug 1 is in fact superior to Drug 2, then choosing action \( a = 1 \) will incur a smaller loss than choosing \( a = 2 \). Now, the usual problem is that we do not know the true state of nature, we only have data that lets us make probabilistic statements about it (ie, we have a posterior distribution for \( \theta \), but do not usually know the exact value of \( \theta \)). Also, we rarely make decisions before seeing the data, so that in general, \( a = a(x) \) is a function of the data. Note that while we will refer to these as “losses”, we could equally well use “gains”.

8. Expected Bayes Loss (Bayes Risk): We do not know the true value of \( \theta \), but we do have a posterior distribution once the data are known, \( f(\theta| x) \). Hence, to make a “coherent” Bayesian decision, we minimize the Expected Bayesian Loss, defined by:

\[
\text{EBL} = \int L(\theta, a(x)) f(\theta| x) d\theta
\]

In other words, we choose the action \( a(x) \) such that the EBL is minimized.

The first five elements in the above list comprise a non-decision theoretic Bayesian approach to statistical inference. This type of analysis (ie, non-decision theoretic) is what most of us are used to seeing in the medical literature. However, many Bayesians argue that the main reason we carry out any statistical analyses is to help in making decisions, so that elements 6, 7, and 8 are crucial. There is little doubt that we will see more such analyses in the near future, but it remains to be seen how popular the decision theoretic framework will become in medicine. The main problem is to specify the loss functions, since there are so many possible consequences (main outcomes, side-effects, costs, etc.) to medical decisions, and it is difficult to combine these into a single loss function. My guess is that much work will have to be done on developing loss functions before the decision theoretic approach becomes mainstream. This course, therefore, will focus on elements 1 through 5.
**Simple Univariate Inference for Common Situations**

As you may have seen in your previous classes and in your experience, many data analyses begin with very simple univariate analyses, using models such as the normal (for continuous data), the binomial (for dichotomous data), the Poisson (for count data) and the multinomial (for multicategorical data).

Here we will see how analyses typically proceed for these simple models from a Bayesian viewpoint.

As described above, in Bayesian analyses, aside from a data model (by which I mean the likelihood function), we need a prior distribution over all unknown parameters in the model. Thus, here we consider “standard” likelihood-prior combinations for these simple situations.

To begin, here is a summary chart of what we will see:
<table>
<thead>
<tr>
<th>Data Type (summary)</th>
<th>Model (likelihood function)</th>
<th>Conjugate Prior</th>
<th>Posterior Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous ($x, n$)</td>
<td>$\text{normal}(\mu, \sigma^2)$</td>
<td>$\text{normal}(\theta, \tau^2)$</td>
<td>$\text{normal}\left(\frac{\mu}{\frac{1}{\sigma^2} + \frac{n}{\tau^2} x}, \left[\frac{1}{\sigma^2} + \frac{n}{\tau^2} \right]^{-1}\right)$</td>
</tr>
<tr>
<td>Dichotomous ($x, n$)</td>
<td>$\text{binomial}(\theta, n)$</td>
<td>$\text{beta}(\alpha, \beta)$</td>
<td>$\text{beta}(\alpha + x, \beta + (n - x))$</td>
</tr>
<tr>
<td>Count ($x$)</td>
<td>$\text{Poisson}(\lambda)$</td>
<td>$\text{gamma}(\alpha, \beta)$</td>
<td>$\text{gamma}(\alpha + x, \beta + 1)$</td>
</tr>
<tr>
<td>Multicat ($x_1, x_2, \ldots, x_m$)</td>
<td>$\text{multinom}(p_1, p_2, \ldots, p_m)$</td>
<td>$\text{dirich}(\alpha_1, \alpha_2, \ldots, \alpha_m)$</td>
<td>$\text{dirich}(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_p + x_m)$</td>
</tr>
</tbody>
</table>
We will now look at each of these four cases in detail.

**Bayesian Inference For A Single Normal Mean**

**EXAMPLE:** Consider the situation where we are trying to estimate the mean diastolic blood pressure of Americans living in the United States from a sample of 27 patients. The data are:

76, 71, 82, 63, 76, 64, 74, 70, 64, 75, 78, 66, 62, 79, 82, 78, 62, 72, 83, 79, 41, 80, 77, 67.

[Note: These are in fact real data obtained from an experiment designed to estimate the effects of calcium supplementation on blood pressure. These are the baseline data for 27 subjects from the study, whose reference is: Lyle, R.M., Melby, C.L., Hyner, G.C., Edmonson, J.W., Miller, J.Z., and Weinberger, M.H. (1987). Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men. *Journal of the American Medical Association*, **257**, 1772–1776.]

From this data, we find $\bar{x} = 71.89$, and $s^2 = 85.18$, so that $s = \sqrt{85.18} = 9.22$

Let us assume the following:

1. The standard deviation is known *a priori* to be 9 mm Hg.
2. The observations come from a Normal distribution, i.e.,

\[ x_i \sim N(\mu, \sigma^2 = 9^2), \quad \text{for } i = 1, 2, \ldots, 27. \]

We will follow the three usual steps used in Bayesian analyses:

1. Write down the likelihood function for the data.
2. Write down the prior distribution for the unknown parameter, in this case $\mu$.
3. Use Bayes theorem to derive the posterior distribution. Use this posterior distribution, or summaries of it like 95% credible intervals for statistical inferences.
**Step 1:** The likelihood function for the data is based on the Normal distribution, i.e.,

\[
f(x_1, x_2, \ldots, x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \sigma}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi \sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right).
\]

**Step 2:** Suppose that we have a priori information that the random parameter \( \mu \) is likely to be in the interval \((60,80)\). That is, we think that the mean diastolic blood pressure should be about 70, but would not be too surprised if it were as low as perhaps 60, or as high as about 80. We will represent this prior distribution as a second Normal distribution (not to be confused with the fact that the data are also assumed to follow a Normal density). The Normal prior density is chosen here for the same reason as the Beta distribution is chosen when we looked at the binomial distribution: it makes the solution of Bayes Theorem very easy. We can therefore approximate our prior knowledge as:

\[
\mu \sim N(\theta, \tau^2) = N(70, 5^2 = 25).
\]

In general, this choice for a prior is based on any information that may be available at the time of the experiment. In this case, the prior distribution was chosen to have a somewhat large standard deviation \( \tau = 5 \) to reflect that we have very little expertise in blood pressures of average Americans. A clinician with experience in this area may elect to choose a much smaller value for \( \tau \). The prior is centered around \( \mu = 70 \), our best guess.

We now wish to combine this prior density with the information in the data to derive the posterior distribution. This combination is again carried out by a version of Bayes Theorem.

\[
\text{posterior distribution} = \frac{\text{prior distribution} \times \text{likelihood of the data}}{\text{a normalizing constant}}
\]

The precise formula is

\[
f(\mu | x_1, \ldots, x_n) = \frac{f(\mu) \times f(x_1, \ldots, x_n | \mu)}{\int_{-\infty}^{+\infty} f(\mu) \times f(x_1, \ldots, x_n | \mu) \, d\mu}
\]
In our case, the prior is given by the Normal density discussed above, and the likelihood function was the product of Normal densities given in Step 1.

Using Bayes Theorem, we multiply the likelihood by the prior, so that after some algebra, the posterior distribution is given by:

\[
\text{Posterior of } \mu \sim N \left( A \times \theta + B \times \bar{x}, \frac{\tau^2 \sigma^2}{n \tau^2 + \sigma^2} \right)
\]

where

\[A = \frac{\sigma^2}{\tau^2 + \sigma^2 / n} = 0.107\]
\[B = \frac{\tau^2}{\tau^2 + \sigma^2 / n} = .893\]
\[n = 27\]
\[\sigma = 9\]
\[\tau = \sqrt{25} = 5\]
\[\theta = 70, \text{ and }\]
\[\bar{x} = 71.89\]

Hence \( \mu \sim N(71.69, 2.68) \), so that graphically, the prior and posterior distributions are:
The mean value depends on both the prior mean, $\theta$, and the observed mean, $\bar{x}$.

Again, the posterior distribution is interpreted as the actual probability density of $\mu$ given the prior information and the data, so that we can calculate the probabilities of being in any interval we like. These calculations can be done in the usual way, using normal tables or R. For example, a 95% credible interval is given by (68.5, 74.9).

**Bayesian Inference For Binomial Proportion**

Suppose that in a given experiment $x$ successes are observed in $N$ independent Bernoulli trials. Let $\theta$ denote the true but unknown probability of success, and suppose that the problem is to find an interval that covers the most likely
locations for \( \theta \) given the data.

The Bayesian solution to this problem follows the usual pattern, as outlined earlier. Here we consider only the first five steps, so that we ignore the decision analysis aspects. Hence the steps of interest can be summarized as:

1. Write down the likelihood function for the data.
2. Write down the prior distribution for the data.
3. Use Bayes theorem to derive the posterior distribution. Use this posterior distribution, or summaries of it like 95% credible intervals for statistical inferences.

For the case of a single binomial parameter, these steps are realized by:

1. The likelihood is the usual binomial probability formula, the same one used in frequentist analysis,

\[
L(x|\theta) = Pr\{x \text{ successes in } N \text{ trials}\} = \frac{N!}{(N-x)!x!} \theta^x (1 - \theta)^{(N-x)}.
\]

In fact, all one needs to specify is that

\[
L(x|\theta) = Pr\{x \text{ successes in } N \text{ trials}\} \propto \theta^x (1 - \theta)^{(N-x)},
\]

since \( \frac{N!}{(N-x)!x!} \) is simply a constant that does not involve \( \theta \). In other words, inference will be the same whether one uses this constant or ignores it.

2. Although any prior distribution can be used, a convenient prior family is the Beta family, since it is the conjugate prior distribution for a binomial experiment. A random variable, \( \theta \), has a distribution that belongs to the Beta family if it has a probability density given by

\[
f(\theta) = \begin{cases} \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, & 0 \leq \theta \leq 1, \ \alpha, \beta > 0, \\ 0, & \text{otherwise,} \end{cases}
\]

[ \( B(\alpha, \beta) \) represents the Beta function evaluated at \((\alpha, \beta)\). It is simply the normalizing constant that is necessary to make the density integrate
to one, that is, $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}dx$. The mean of the Beta distribution is given by
\[
\mu = \frac{\alpha}{\alpha + \beta},
\]
and the standard deviation is given by
\[
\sigma = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}}.
\]
Therefore, at this step, one needs only to specify $\alpha$ and $\beta$ values, which can be done by finding the $\alpha$ and $\beta$ values that give the correct prior mean and standard deviation values. This involves solving two equations in two unknowns. The solution is
\[
\alpha = -\frac{\mu (\sigma^2 + \mu^2 - \mu)}{\sigma^2}
\]
and
\[
\beta = \frac{(\mu - 1) (\sigma^2 + \mu^2 - \mu)}{\sigma^2}
\]
3. As always, Bayes Theorem says
\[
\text{posterior distribution } \propto \text{ prior distribution } \times \text{ likelihood function.}
\]
In this case, it can be shown (by relatively simple algebra) that if the prior distribution is $Beta(\alpha, \beta)$, and the data is $x$ successes in $N$ trials, then the posterior distribution is $Beta(\alpha + x, \beta + N - x)$.

**Example:** Suppose that a new diagnostic test for a certain disease is being investigated. Suppose that 100 persons with confirmed disease are tested, and that 80 of these persons test positively.

(a) What is the posterior distribution of the sensitivity of the test if a Uniform $Beta(\alpha = 1, \beta = 1)$ prior is used? What is the posterior mean and standard deviation of this distribution?

(b) What is the posterior distribution of the sensitivity of the test if a $Beta(\alpha = 27, \beta = 3)$ prior is used? What is the posterior mean and standard deviation of this distribution?

(c) Draw a sketch of the prior and posterior distributions from both (a) and (b).
(d) Derive the 95% posterior credible intervals from the two posterior distributions given above, and compare it to the usual frequentist confidence interval for the data. Clearly distinguish the two different interpretations given to confidence intervals and credible intervals.

Solution:

(a) According to the result given above, the posterior distribution is again a Beta, with parameters $\alpha = 1 + 80 = 81$, $\beta = 1 + 20 = 21$. The mean of this distribution is $\frac{81}{81 + 21} = 0.794$, and the standard deviation is 0.0398.

(b) Again the posterior distribution is a Beta, with parameters $\alpha = 27 + 80 = 107$, $\beta = 3 + 20 = 23$. The mean of this distribution is $\frac{107}{107 + 23} = 0.823$, and the standard deviation is 0.0333.

(c) See Below.
(d) From tables of the beta density (contained in many books of statistical
tables) or software that includes Bayesian analysis, the 95% credible intervals
are (0.71, 0.86) from the Beta(81,21) posterior density, and (0.75, 0.88) from
the Beta(107,23) posterior density. The frequentist 95% confidence interval is
(0.71, 0.87).

Note that numerically, the frequentist confidence interval is nearly identical to
the Bayesian credible interval starting from a Uniform prior. However, their
interpretations are very different. Credible intervals are interpreted directly
as the posterior probability that \( \theta \) is in the interval, given the data and the prior
distribution. No references to long run frequencies or other experiments are
required. On the other hand, confidence intervals have the interpretation that
if such procedures are used repeatedly, then 100(1 – \( \alpha \))% of all such sets would
in the long run contain the true parameter of interest. Notice that there can be
nothing said about what happened in this particular case, the only inference is
to the long run. To infer anything about the particular case from a frequentist
analysis involves a “leap of faith.”

**Bayesian Inference For Multinomial Proportions**

Recall that the multinomial distribution, given by

\[
f(x_1, x_2, \ldots, x_m; p_1, p_2, \ldots, p_m) = \binom{n}{x_1, x_2, \ldots, x_m} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}
\]

is simply a multivariate version of the binomial distribution. While the bi-
nomial accommodates only two possible outcomes (yes/no, success/failure,
males/females, etc.), the multinomial allows for analyses of categorical data with
more than two categories (yes/no/maybes, Liberal/Conservative/NDP, items
rated on a 1 to 5 scale, etc.).

Similarly, the Dirichlet distribution is simply a multivariate version of the
beta density, given by

\[
f(p_1, p_2, \ldots, p_m; \alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_m)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \times \cdots \times \Gamma(\alpha_m)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \times \cdots \times p_m^{\alpha_m-1}
\]
All two-dimensional marginal densities of Dirichlet distributions are beta densities. The mean probability in the $i^{th}$ category (similar again to the beta) is

$$\frac{\alpha_i}{\sum_{j=1}^{m} \alpha_j}.$$

As indicated in the chart, similar to the result for a beta distribution, if we begin with a prior distribution that is Dirichlet, and we have multinomial data, our posterior distribution is also a Dirichlet distribution. As indicated in the chart, and similar to the case of a beta-binomial model, in the Dirichlet-multinomial model, again, we can simply add the prior and data values together, category by category, to derive the final posterior distribution.

Example: Suppose we track slightly overweight Canadians for five years, to see if they remain slightly overweight, become more overweight (obese), or become normal weight (these categories are usually defined by BMI values, i.e., body mass index). Suppose out of 100 slightly overweight persons tracked, 75 remain slightly overweight, 10 become obese, and 15 become normal weight. Suppose you believe that about 50% would stay in their original weight category, and half of the rest would move to each adjacent category. Suppose also that you believe your prior information to be the equivalent of 10 observations. Provide the prior and posterior distributions. Also, provide the marginal distribution for the probability of remaining slightly overweight.

Solution: Prior would be Dirichlet(2.5, 5, 2.5). Posterior would be Dirichlet(2.5 + 15, 5 + 75, 2.5 + 10) = Dirichlet(17.5, 80, 12.5). Marginal posterior for middle category is beta(80, 30).

Bayesian Inference For Poisson Count Data

Recall that the Poisson distribution is given by:

$$p(y|\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

where $y$ is an observed count, and $\lambda$ is the rate of the Poisson distribution. Recall that both the mean and variance of the Poisson distribution is given by the rate $\lambda$. 
Further, recall (see math background section of course) that the Gamma distribution is given by:

\[ f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp(-\beta\lambda)\lambda^{\alpha-1}, \text{ for } \lambda > 0. \]

Although one is a discrete distribution and the other is continuous, note the similarity of form between the Poisson and Gamma distributions: Both have a parameter raised to a power, and both have a simple exponential term. The Gamma can be shown to be a conjugate prior distribution for the Poisson likelihood function.

Suppose count data \( y \) follow a Poisson distribution with parameter \( \lambda \), and the prior distribution put on \( \lambda \) is Gamma(\( \alpha, \beta \)), where \( \alpha \) and \( \beta \) are known constants. Then it can be shown that the posterior distribution for \( \lambda \) is again a Gamma density, in particular Gamma(\( \alpha + y, \beta + 1 \)). The prior parameters \( \alpha \) and \( \beta \) can be thought of as having observed a total count of \( \alpha - 1 \) events in \( \beta \) prior observations.

**Prior Distributions**

Various families of prior distributions can be defined:

**Low Information Priors:** Prior distributions such as the uniform over the range of the variable say that all values in the feasible range are equally likely *a priori*, so that the posterior distribution is determined solely (or almost) by the data. This prior allows one to see what inferences would be available from the current data set alone, ignoring any past information (like all frequentist analyses).

**Clinical Priors:** Summarizes the “average beliefs” across knowledgeable clinicians.

**Optimistic Priors:** Summarize beliefs of clinicians (or others) who are optimistic about a treatment under study.

**Pessimistic Priors:** Summarize beliefs of clinicians (or others) who are pessimistic about a treatment under study.
Most Bayesian analyses will be carried out using at least one choice of low information prior. Then, if warranted, other priors may also be used. If there is a considerable range of prior opinions in the community, posterior distributions from all four of the above priors may be calculated and reported.