Introduction to Bayesian Inference

Mohan Delampady
Statistics and Mathematics Unit
Indian Statistical Institute, Bangalore

July 4, 2007

Outline

What is Statistical Inference?

It is an inverse problem as in 'Toy Example'.

Example 1 (Toy). Suppose a million candidate stars are examined for the presence of planetary systems associated with them. If 272 'successes' are noticed, how likely that the success rate is 1%, 0.1%, 0.01%, $\cdots$ for the entire universe?

Probability models for observed data involve *direct* probabilities as in Example 2.

Example 2. An urn has 100 marbles of which 20 are red and the rest blue. 10 marbles are drawn at random with replacement (repeatedly, one by one, after replacing the one previously drawn and mixing the marbles well). How many marbles drawn will be red?

Data and Models

X = number of red marbles in the sample (out of sample size $n = 10$)

$$P(X = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{(n-k)}, \quad k = 0, 1, \ldots n \qquad (1)$$

In (1) θ is the proportion of red marbles in the urn, which is also the probability of drawing a red marble at each draw. In Example 2, $\theta = \frac{20}{100} = 0.2$ and $n = 10$. So,

$P(X = 0|\theta = 0.2) = 0.8^{10}$, $P(X = 1|\theta = 0.2) = 10 \times 0.2 \times 0.8^9$, and so on.

In practice, as in 'Toy Example', θ is unknown and inference about it is the question to solve.

In the Urn example, if θ is not known and 3 marbles out of 10 turned out to be red, one could ask:

how likely is $\theta = 0.1$, or 0.2 or 0.3 or ...?

Thus inference about θ is an inverse problem:

Causes (parameters) $\longleftarrow$ Effects (observations)

How does this *inversion* work?

The direct probability model $P(X = k|\theta)$ provides a *likelihood function* for the unknown *parameter* θ when data $X = x$ is observed:

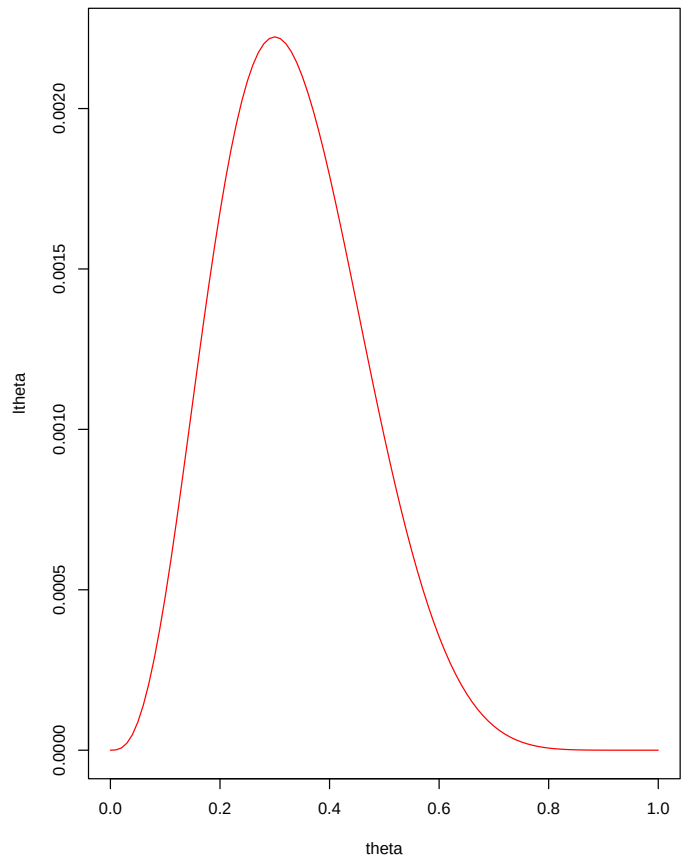
$l(\theta|x) = f(x|\theta)$ $(= P(X = x|\theta)$ when X is a discrete random variable) as function of θ for given x.

Interpretation: $f(x|\theta)$ says how likely x is under different θ or the model $P(.|\theta)$, so if x is observed, then
$P(X = x|\theta) = f(x|\theta) = l(\theta|x)$ should be able to indicate what the likelihood of different θ values or $P(.|\theta)$ are for that x.

As a function of x for fixed θ $P(X = x|\theta)$ is a probability mass function or density, but as a function of θ for fixed x, it has no such meaning, but just a measure of likelihood.

After an experiment is conducted and seeing data x, the only entity available to convey the information about θ obtained from the experiment is $l(\theta|x)$.

For the Urn Example we have $l(\theta|X = 3) \propto \theta^3(1-\theta)^7$:



Maximum Likelihood Estimation (MLE): If $l(\theta|x)$ measures the likelihood of different θ (or the corresponding models $P(.|\theta)$), just find that $\theta = \hat{\theta}$ which maximizes the likelihood.

For model (1)

$$\hat{\theta} = \hat{\theta}(x) = x/n = \text{ sample proportion of successes .}$$

This is only an estimate. How good is it? What is the possible error in estimation?

Likelihood function $l(\theta|x)$ has nothing to say about these.

Frequentist Statistics

Consider repeating this experiment again and again. Then one can look at all possible sample data. i.e. all possible x values. Utilize *long-run average behaviour* of the MLE. i.e. treat $\hat{\theta}$ as a random quantity by replacing x by X in $\hat{\theta}(x)$. i.e. look at X/n where X can take all possible values, 0, 1, ... n.

$X \sim$ Binomial(n, θ) with the probability model (1). Noting that the variance of such an X is $n\theta(1-\theta)$, one obtains the variance of X/n to be $\theta(1-\theta)/n$, which can be estimated by $\hat{\theta}(1-\hat{\theta})/n$. A measure of estimation error of $\hat{\theta}$ is the estimated standard deviation of X/n, namely, $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. For further development we need large n, so that we can apply the *Law of Large Numbers* and the *Central Limit Theorem* to X/n. Then, the estimator will be close to the true θ probabilistically and also, it is approximately distributed like a Gaussian random variable with mean θ and variance $\theta(1-\theta)/n$.

Confidence Statements

Specifically, for large n, approximately

$$\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} \sim N(0,1),$$

or

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \sim N(0,1). \tag{2}$$

From (2), a large n and approximate 95% confidence interval for θ is

$$\hat{\theta} \pm 2\sqrt{\hat{\theta}(1-\hat{\theta})/n}.$$

What Does This Mean?

Simply, if we sample again and again, in about 19 cases out of 20 this random interval

$$\left(\hat{\theta}(X) - 2\sqrt{\hat{\theta}(X)(1-\hat{\theta}(X))/n}, \hat{\theta}(X) + 2\sqrt{\hat{\theta}(X)(1-\hat{\theta}(X))/n}\right)$$

will contain the true unknown value of θ.

Fine, but what can we say about the one interval that we can construct for the given sample or data x?

Nothing; either θ is inside $(0.3 - 2\sqrt{0.3 \times 0.7/10}, 0.3 + 2\sqrt{0.3 \times 0.7/10})$ or it is outside.

Can we say $0.3 - 2\sqrt{0.3 \times 0.7/10} \leq \theta \leq 0.3 + 2\sqrt{0.3 \times 0.7/10}$ with 95% chance?

Not in this approach. If θ is treated as fixed unknown constant, conditioning on the given data $X = x$ is meaningless.

Conditioning on Data

- What other approach is possible, then?
- How does one condition on data?
- How does one talk about probability of a model or a hypothesis?

Example 3.(not from physics but medicine) Consider a blood test for a certain disease; result is *positive* ($x = 1$) or *negative* ($x = 0$). Suppose θ_1 denotes *disease is present*, θ_2 *disease not present*.

Test is not confirmatory. Instead the probability distribution of X for different θ is:

	$x = 0$	$x = 1$	What does it say?
θ_1	0.2	0.8	Test +ve 80% of time if 'disease present'
θ_2	0.7	0.3	Test −ve 70% of time if 'disease not present'

If for a particular patient the test result comes out to be 'positive', what should the doctor conclude?

What is the Question?

What is to be answered is 'what are the chances that the *disease is present* given that the test is positive?' i.e., $P(\theta = \theta_1 | X = 1)$.

What we have is $P(X = 1 | \theta = \theta_1)$ and $P(X = 1 | \theta = \theta_2)$.

We have the 'wrong' conditional probabilities. They need to be 'reversed'. But how?

The Bayesian Recipe

Recall the Bayes Theorem: If A and B are two events,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

assuming $P(B) > 0$. Therefore, $P(A \text{ and } B) = P(A|B)P(B)$, and by symmetry $P(A \text{ and } B) = P(B|A)P(A)$. Consequently, if $P(B|A)$ is given and $P(A|B)$ is desired, note

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

Rule of total probability says,

$$\begin{aligned} P(B) = P(B \text{ and } \Omega) &= P(B \text{ and } A) + P(B \text{ and } A^c) \\ &= P(B|A)P(A) + P(B|A^c)(1 - P(A)), \text{ so} \end{aligned}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)(1 - P(A))} \qquad (3)$$

Bayes Theorem allows one to invert a certain conditional probability to get a certain other conditional probability. How does this help us?

In our example we want $P(\theta = \theta_1 | X = 1)$. From (3),

$$P(\theta = \theta_1 \mid X = 1) = \frac{P(X = 1 \mid \theta = \theta_1)P(\theta = \theta_1)}{P(X = 1 \mid \theta = \theta_1)P(\theta = \theta_1) + P(X = 1 \mid \theta = \theta_2)P(\theta = \theta_2)} \quad (4)$$

So, all we need is $P(\theta = \theta_1)$, which is simply the probability that a randomly chosen person has this disease, or just the 'prevalence' of this disease in the concerned population. The good doctor most likely has this information from his experience in the field. But this is not part of the experimental data. This is pre-experimental information or *prior* information. If we have this, and are willing to incorporate it in the analysis, we get the post-experimental information or *posterior* information in the form of $P(\theta | X = x)$.

In our example, if we take $P(\theta = \theta_1) = 0.05$ or 5%, we get

$$P(\theta = \theta_1 \mid X = 1) = \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.3 \times 0.95} = \frac{0.04}{0.325} = 0.123$$

which is only 12.3% and $P(\theta = \theta_2 \mid X = 1) = 0.877$ or 87.7%.

Formula (4) which shows how to 'invert' the given conditional probabilities, $P(X = x \mid \theta)$ into the conditional probabilities of interest, $P(\theta \mid X = x)$ is an instance of the Bayes Theorem, and hence the *Theory of Inverse Probability* (usage at the time of Bayes and Laplace, late eighteenth century and even by Jeffreys), is known these days as *Bayesian inference*.

Ingredients of Bayesian inference:

likelihood function, $l(\theta|x)$; θ can be a parameter vector

prior probability, $\pi(\theta)$

Combining the two, one gets the posterior probability density or mass function

$$\pi(\theta \mid x) = \begin{cases} \frac{\pi(\theta) l(\theta|x)}{\sum_j \pi(\theta_j) l(\theta_j|x)} & \text{if } \theta \text{ is discrete;} \\ \frac{\pi(\theta) l(\theta|x)}{\int \pi(u) l(u|x)\, du} & \text{if } \theta \text{ is continuous.} \end{cases} \tag{5}$$

Inference for Binomial proportion

Example 2 contd. Suppose we have no special information available on θ. Then assume θ is uniformly distributed on the interval $(0, 1)$. i.e., the prior density is $\pi(\theta) = 1, \quad 0 < \theta < 1$.

This is a choice of *non-informative* or *vague* or *reference* prior. Often, Bayesian inference from such a prior coincides with classical inference.

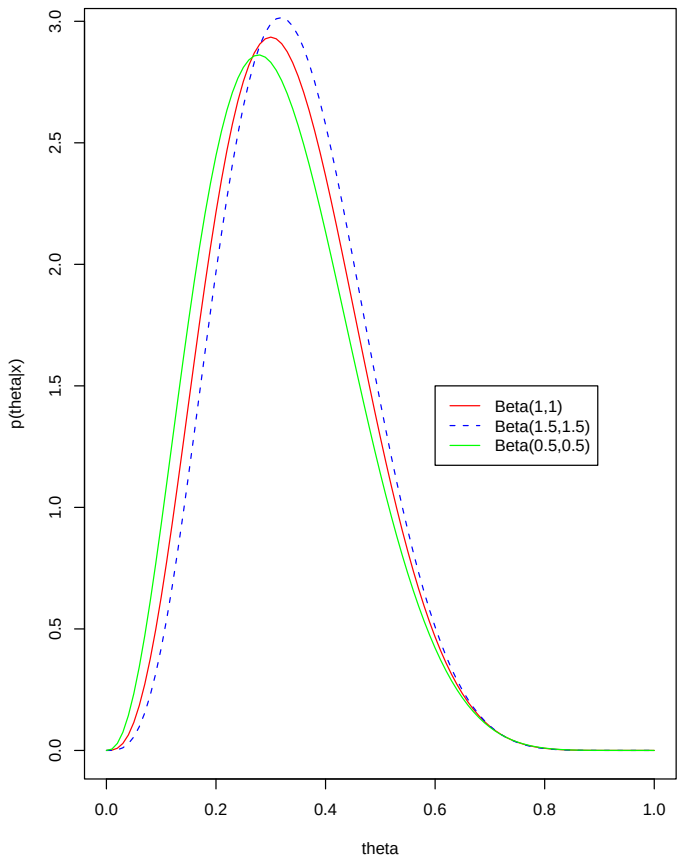
In the Example then the posterior density of θ given x is

$$\begin{aligned} \pi(\theta|x) &= \frac{\pi(\theta)l(\theta|x)}{\int \pi(u)l(u|x)\,du} \\ &= \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, \quad 0 < \theta < 1. \end{aligned}$$

As functions of θ, this is the same as the likelihood function $l(\theta|x) \propto \theta^x(1-\theta)^{n-x}$, and so maximizing the posterior probability density will give the same estimate as the maximum likelihood estimate!

Influence of the Prior

If we had some knowledge of θ which can be summarized in the form of a Beta prior distribution with parameters α and γ, the posterior will be also Beta with parameters $x + \alpha$ and $n - x + \gamma$. Such priors which result in posteriors from the same 'family' are called 'natural conjugate priors'. Robustness?



Robustness?

Objective Bayesian Analysis:

Invariant priors: Jeffreys

Reference priors: Bernardo, Jeffreys

Maximum entropy priors: Jaynes

In Example 2, what $\pi(\theta|x)$ says is that the uncertainty in θ can now be described in terms of an actual probability distribution concentrated around the maximum likelihood estimate $\hat{\theta} = x/n$. However, the interpretation of $\hat{\theta}$ as an estimate of θ is quite different. It is the most probable value of the unknown parameter θ conditional on the sample data x; it is called the 'maximum a posteriori estimate (MAP)' or the 'highest posterior density estimate (HPD)'.

There is no need to mimic the MLE anymore. We have a genuine probability distribution, namely, the posterior distribution to quantify our post-experimental knowledge about θ. Indeed the usual Bayes estimate is the mean of the posterior distribution which minimizes the posterior dispersion:

$$E[(\theta - \hat{\theta}_B)^2|x] = \min_a E[(\theta - a)^2|x],$$

when $\hat{\theta}_B = E(\theta|x)$.

If we choose $\hat{\theta}_B$ as the estimate of θ, we get a natural measure of variability of this estimate in the form of the posterior variance: $E[(\theta - E(\theta|x))^2|x]$. Therefore the posterior standard deviation is a natural measure of estimation error. i.e., our estimate is $\hat{\theta}_B \pm \sqrt{E[(\theta - E(\theta|x))^2|x]}$.

In fact, we can say much more. For any interval around $\hat{\theta}$ we can compute the (posterior) probability of it containing the true parameter θ. In other words, a statement such as

$$P(\hat{\theta}_B - k_1 \leq \theta \leq \hat{\theta}_B + k_2|x) = 0.95$$

is perfectly meaningful.

All these inferences are conditional on the given data.

In Example 2, if the prior is a Beta distribution with parameters α and γ, then $\theta|x$ will have a Beta$(x+\alpha, n-x+\gamma)$ distribution, so the Bayes estimate of θ will be

$$\hat{\theta}_B = \frac{(x+\alpha)}{(n+\alpha+\gamma)} = \frac{n}{n+\alpha+\gamma}\frac{x}{n} + \frac{\alpha+\gamma}{n+\alpha+\gamma}\frac{\alpha}{\alpha+\gamma}.$$

This is a convex combination of sample mean and prior mean, with the weights depending upon the sample size and the strength of the prior information as measured by the values of α and γ.

Bayesian inference relies on the conditional probability language to revise one's knowledge. In the above example, prior to the collection of sample data one had some (vague, perhaps) information on θ. Then came the sample data. Combining the model density of this data with the prior density one gets the posterior density, the conditional density of θ given the data. From now on until further data is available, this posterior distribution of θ is the only relevant information as far as θ is concerned.

Inference With Normals/Gaussians

Gaussian PDF

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{over } [-\infty,\infty] \tag{6}$$

Common abbreviated notation: $X \sim N(\mu,\sigma^2)$

Parameters

$$\mu = E(X) \equiv \langle X \rangle \equiv \int x\, f(x|\mu,\sigma^2)\, dx$$

$$\sigma^2 = E(X-\mu)^2 \equiv \langle (X-\mu)^2 \rangle \equiv \int (x-\mu)^2\, f(x|\mu,\sigma^2)\, dx$$

Inference About a Normal Mean

Example 4. Fit a normal/Gaussian model to the 'globular cluster luminosity functions' data. The set-up is as follows.

Our data consist of n measurements, $X_i = \mu + \epsilon_i$.
Suppose the noise contributions are independent, and $\epsilon_i \sim N(0, \sigma^2)$. Denoting by $\mathbf{x}$, the random sample $(x_1, \ldots x_n)$,

$$\begin{aligned} f(\mathbf{x}|\mu,\sigma^2) &= \prod_i f(x_i|\mu,\sigma^2) \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]}. \end{aligned}$$

Note $(\bar{X}, s^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1))$ is *sufficient* for the parameters (μ, σ^2). This is a very substantial data compression.

Inference About a Normal Mean, σ^2 known

(Not useful, but easy to understand.)

$$l(\mu|\mathbf{x}) \propto f(\mathbf{x}|\mu,\sigma^2) \propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2},$$

so that $\bar{X}$ is sufficient. Also, $\bar{X}|\mu \sim N(\mu, \sigma^2/n)$. If an informative prior, $\mu \sim N(\mu_0, \tau^2)$ is chosen for μ,

$$\begin{aligned}
\pi(\mu|\mathbf{x}) &\propto l(\mu|\mathbf{x})\pi(\mu) \\
&\propto e^{-\frac{1}{2}\left[\frac{n(\mu-\bar{x})^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\tau^2}\right]} \\
&\propto e^{-\frac{\tau^2+\sigma^2/n}{2\tau^2\sigma^2/n}\left(\mu-\frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}(\frac{\mu_0}{\tau^2}+\frac{n\bar{x}}{\sigma^2})\right)^2}.
\end{aligned}$$

i.e., $\mu|\mathbf{x} \sim N(\hat{\mu}, \delta^2)$:

$$\begin{aligned}
\hat{\mu} &= \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}(\frac{\mu_0}{\tau^2}+\frac{n\bar{x}}{\sigma^2}) \\
&= \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu_0.
\end{aligned}$$

$\hat{\mu}$ is the Bayes estimate of μs, which is just a weighted average of sample mean $\bar{x}$ and prior mean μ_0.

δ^2 is the posterior variance of μ and

$$\delta^2 = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n} = \frac{\sigma^2}{n}\frac{\tau^2}{\tau^2 + \sigma^2/n}.$$

Therefore $\hat{\mu} \pm \delta$ is our estimate for μ and $\hat{\mu} \pm 2\delta$ is a 95% HPD (Bayesian) credible interval for μ.

What happens as $\tau^2 \to \infty$, or as the prior becomes more and more flat?

$$\hat{\mu} \to \bar{x}, \quad \delta \to \frac{\sigma}{\sqrt{n}}$$

i.e., Jeffreys' prior $\pi(\mu) = C$ reproduces frequentist inference.

Inference About a Normal Mean, σ^2 unknown

Our observations $X_1, \ldots X_n$ is a random sample from a Gaussian population with both mean μ and variance σ^2 unknown.
We are only interested in μ.
How do we get rid of the nuisance parameter σ^2?
Bayesian inference uses posterior distribution which is a probability distribution, so σ^2 should be integrated out from the joint posterior distribution of μ and σ^2.

$$l(\mu, \sigma^2|\mathbf{x}) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\mu-\bar{x})^2\right]}.$$

Start with $\pi(\mu, \sigma^2)$ and get

$$\pi(\mu, \sigma^2|\mathbf{x}) \propto \pi(\mu, \sigma^2) l(\mu, \sigma^2|\mathbf{x})$$

and then get

$$\pi(\mu|\mathbf{x}) = \int_0^\infty \pi(\mu, \sigma^2|\mathbf{x})\, d\sigma^2.$$

Use Jeffreys' prior $\pi(\mu, \sigma^2) \propto 1/\sigma^2$: Flat prior for μ which is a location or translation parameter, and an independent flat prior for $\log(\sigma)$ which is again a location parameter, being the log of a scale parameter.

$$\pi(\mu, \sigma^2|\mathbf{x}) \propto \frac{1}{\sigma^2} l(\mu, \sigma^2|\mathbf{x})$$

$$\begin{aligned}
\pi(\mu|\mathbf{x}) &\propto \int_0^\infty (\sigma^2)^{-(n+1)/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\bar{x})^2 + n(\mu-\bar{x})^2\right]} d\sigma^2 \\
&\propto \left[(n-1)s^2 + n(\mu-\bar{x})^2\right]^{-n/2} \\
&\propto \left[1 + \frac{1}{n-1}\frac{n(\mu-\bar{x})^2}{s^2}\right]^{-n/2} \\
&\propto \text{ density of Students } t_{n-1}.
\end{aligned}$$

$$\frac{\sqrt{n}(\mu - \bar{x})}{s} | \text{ data} \sim t_{n-1}$$

$$P(\bar{x} - t_{n-1}(0.975)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1}(0.975)\frac{s}{\sqrt{n}} \mid \text{ data}) = 95\%$$

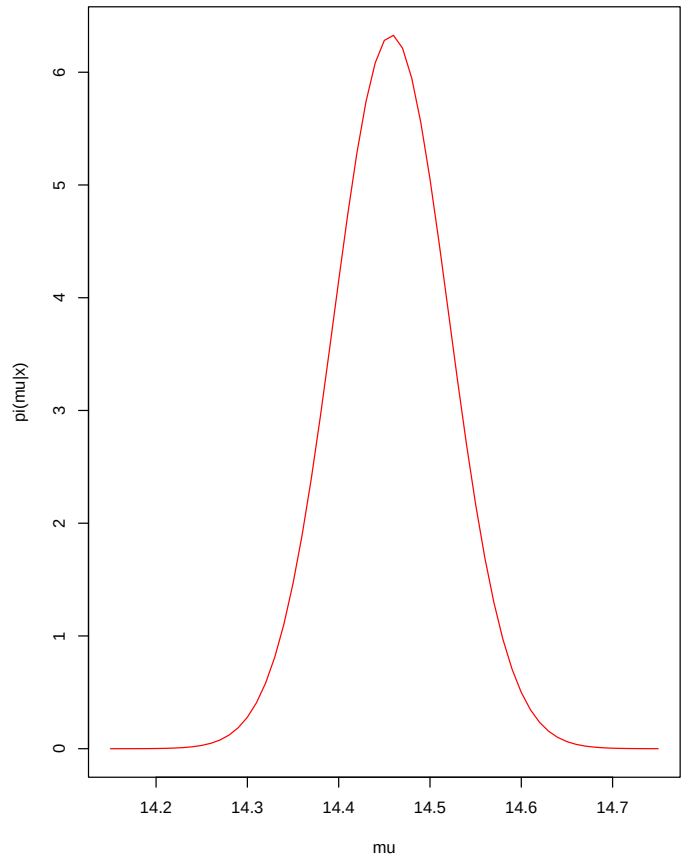
i.e., the Jeffreys' translation-scale invariant prior reproduces frequentist inference.

What if there are some constraints on μ such as $-A \leq \mu \leq B$, for example, $\mu > 0$? We will get a truncated t_{n-1} instead, but the procedure will go through with minimal change.

Example 4 contd. (GCL Data) $n = 360$, $\bar{x} = 14.46$, $s = 1.19$.

$$\frac{\sqrt{360}(\mu - 14.46)}{1.19} | \text{ data} \sim t_{359}$$

$$\mu | \text{ data} \sim N(14.46, 0.063^2) \text{ approximately.}$$



Estimate for mean GCL is 14.46 ± 0.063 and 95% HPD credible interval is $(14.33, 14.59)$.

Comparing two Normal Means

Example 5. Check whether the mean distance indicators in the two populations of LMC datasets are different. Model as follows:

$X_1, \ldots X_{n_1}$ is a random sample from $N(\mu_1, \sigma_1^2)$. $Y_1, \ldots Y_{n_2}$ is a random sample from $N(\mu_2, \sigma_2^2)$. These two samples are independent. Unknown parameters are $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, but quantity of interest is $\eta = \mu_1 - \mu_2$. σ_1^2 and σ_2^2 are nuisance parameters.

Case 1. $\sigma_1^2 = \sigma_2^2$. Check that $\left(\bar{X}, \bar{Y}, s^2 = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2\right)\right)$ is sufficient for (μ_1, μ_2, σ^2). Also, $\bar{X}|\mu_1, \mu_2, \sigma^2 \sim N(\mu_1, \sigma^2/n_1)$, $\bar{Y}|\mu_1, \mu_2, \sigma^2 \sim N(\mu_2, \sigma^2/n_2)$, $(n_1 + n_2 - 2)s^2|\mu_1, \mu_2, \sigma^2 \sim \sigma^2\chi^2_{n_1+n_2-2}$, and these three are independently distributed.

Note $\bar{X} - \bar{Y} | \mu_1, \mu_2, \sigma^2 \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$. Use Jeffreys' location-scale invariant prior $\pi(\mu_1, \mu_2, \sigma^2) \propto 1/\sigma^2$. Noting $\eta = \mu_1 - \mu_2$,

$$\eta | \sigma^2, \mathbf{x}, \mathbf{y} \sim N(\bar{x} - \bar{y}, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})), \text{ and}$$

$$\pi(\eta, \sigma^2 | \mathbf{x}, \mathbf{y}) \propto \pi(\eta | \sigma^2, \mathbf{x}, \mathbf{y}) \pi(\sigma^2 | s^2), \tag{7}$$

integrate out σ^2 from (7) as in the previous example to get

$$\frac{\eta - (\bar{x} - \bar{y})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \mid \mathbf{x}, \mathbf{y} \sim t_{n_1+n_2-2}.$$

95% HPD credible interval for $\eta = \mu_1 - \mu_2$ is

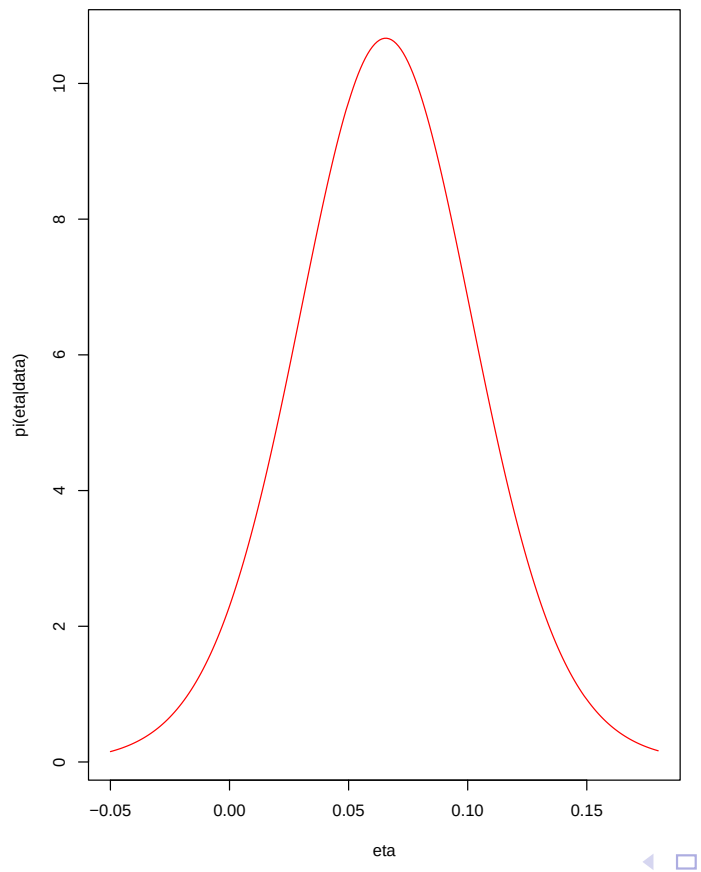
$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2}(0.975) s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

same as frequentist t-interval.

Example 5 contd. We have $\bar{x} = 18.539$, $\bar{y} = 18.473$, $n_1 = 13$, $n_2 = 12$ and $s^2 = 0.0085$. $\hat{\eta} = \bar{x} - \bar{y} = 0.066$, $s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.037$, $t_{23}(0.975) = 2.069$.

95% HPD credible interval for $\eta = \mu_1 - \mu_2$:

$(0.066 - 2.069 \times 0.037, 0.066 + 2.069 \times 0.037) = (-0.011, 0.142)$.



Case 2. σ_1^2 and σ_2^2 are not known to be equal.

From the one-sample normal example, note that
$(\bar{X}, s_X^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2)$ sufficient for (μ_1, σ_1^2), and
$(\bar{Y}, s_Y^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j - \bar{Y})^2)$ sufficient for (μ_2, σ_2^2).

Making inference on $\eta = \mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are not assumed to be equal is called the Behrens-Fisher problem for which the frequentist solution is not very straight forward, but the Bayes solution is.

$\bar{X}|\mu_1, \sigma_1^2 \sim N(\mu_1, \sigma_1^2/n_1)$, $(n_1 - 1)s_X^2|\mu_1, \sigma_1^2 \sim \sigma^2\chi^2_{n_1-1}$, and are independently distributed.

$\bar{Y}|\mu_2, \sigma_2^2 \sim N(\mu_2, \sigma_2^2/n_2)$, $(n_2 - 1)s_Y^2|\mu_2, \sigma_2^2 \sim \sigma^2\chi^2_{n_2-1}$, and are independently distributed.

X and **Y** samples are independent.

Use Jeffreys' prior $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto 1/\sigma_1^2 \times 1/\sigma_2^2$

Calculations similar to those in one-sample case give:

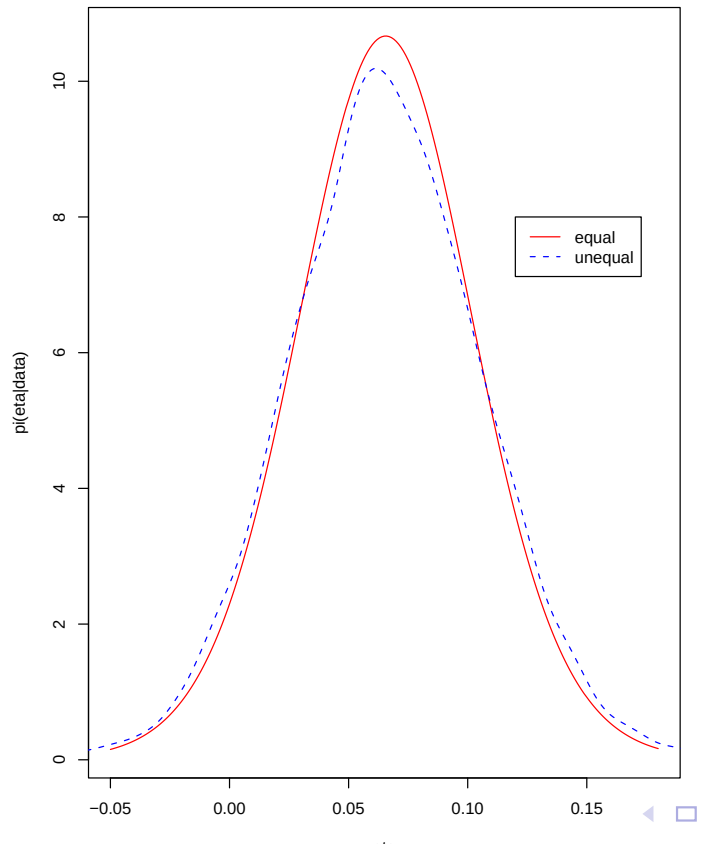
$$\begin{aligned} \frac{\sqrt{n_1}(\mu_1 - \bar{x})}{s_X} | \text{ data} &\sim t_{n_1-1}, \\ \frac{\sqrt{n_2}(\mu_2 - \bar{y})}{s_Y} | \text{ data} &\sim t_{n_2-1}, \end{aligned} \tag{8}$$

and these two are independent.

Posterior distribution of $\eta = \mu_1 - \mu_2$ given the data is non-standard (difference of two independent t variables) but not difficult to obtain.

Use Monte-Carlo Sampling: Simply generate (μ_1, μ_2) repeatedly from (8) and construct a histogram for $\eta = \mu_1 - \mu_2$

Example 5 (LMC) contd. Looks slightly different.



Posterior mean of $\eta = \mu_1 - \mu_2$ is

$$\hat{\eta} = E(\mu_1 - \mu_2 | \text{ data}) = \begin{cases} 0.0656 & \text{equal variance;} \\ 0.0657 & \text{unequal variance.} \end{cases} \tag{9}$$

95% HPD credible interval for $\eta = \mu_1 - \mu_2$ is

$$= \begin{cases} (-0.011, 0.142) & \text{equal variance;} \\ (-0.014, 0.147) & \text{unequal variance.} \end{cases} \tag{10}$$

Empirical Bayes Methods for High Dimensional Problems

This is becoming popular again, this time for 'high dimensional' problems. Astronomers routinely estimate characteristics of millions of similar astronomical objects – distance, radial velocity whatever. Consider the data:

$$(\mathbf{X}_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n} \end{pmatrix}, \mathbf{X}_2 = \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2n} \end{pmatrix}, \cdots \mathbf{X}_p = \begin{pmatrix} X_{p1} \\ X_{p2} \\ \vdots \\ X_{pn} \end{pmatrix}).$$

$\mathbf{X}_j$ represents n repeated independent observations on the jth object, $j = 1, 2, \ldots p$. The important point is n is small, 2, 5, or 10, whereas p is large, such as a million.

Suppose $X_{j1}, \ldots X_{jn}$ measure μ_j with variability σ^2.

Problem: Maximum likelihood can give wrong estimates

Take $n = 2$ and suppose

$$\begin{pmatrix} X_{j1} \\ X_{j2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_j \\ \mu_j \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right), \quad j = 1, 2, \ldots p.$$

i.e., we measure μ_j with 2 independent measurements, each coming with a $N(0, \sigma^2)$ error added to it; we do this for a very large number p of objects. What is the MLE of σ^2?

$$\begin{aligned} l(\mu_1, \ldots \mu_p; \sigma^2 | \mathbf{x}_1, \ldots \mathbf{x}_p) &= f(\mathbf{x}_1, \ldots \mathbf{x}_p | \mu_1, \ldots \mu_p; \sigma^2) \\ &= \prod_{j=1}^{p} \prod_{i=1}^{2} f(x_{ji} | \mu_j, \sigma^2) \\ &= (2\pi\sigma^2)^{-p} \exp(-\frac{1}{2\sigma^2} \sum_{j=1}^{p} \sum_{i=1}^{2} (x_{ji} - \mu_j)^2) \\ &= (2\pi\sigma^2)^{-p} \exp(-\frac{1}{2\sigma^2} \sum_{j=1}^{p} \left[\sum_{i=1}^{2} (x_{ji} - \bar{x}_j)^2 + 2(\bar{x}_j - \mu_j)^2 \right]). \end{aligned}$$

$\hat{\mu}_j = \bar{x}_j = (x_{j1} + x_{j2})/2$ and

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{2p}\sum_{j=1}^{p}\sum_{i=1}^{2}(x_{ji} - \bar{x}_j)^2 \\
&= \frac{1}{2p}\sum_{j=1}^{p}\left[\left(x_{j1} - \frac{x_{j1}+x_{j2}}{2}\right)^2 + \left(x_{j2} - \frac{x_{j1}+x_{j2}}{2}\right)^2\right] \\
&= \frac{1}{2p}\sum_{j=1}^{p} 2\frac{(x_{j1}-x_{j2})^2}{4} = \frac{1}{4p}\sum_{j=1}^{p}(x_{j1}-x_{j2})^2.
\end{aligned}$$

Since $X_{j1} - X_{j2} \sim N(0, 2\sigma^2)$, $j = 1, 2 \ldots,$

$$\frac{1}{p}\sum_{j=1}^{p}(X_{j1} - X_{j2})^2 \xrightarrow[p\to\infty]{P} 2\sigma^2, \text{ so that}$$

$$\hat{\sigma^2} = \frac{1}{4p}\sum_{j=1}^{p}(X_{j1} - X_{j2})^2 \xrightarrow[p\to\infty]{P} \frac{\sigma^2}{2}, \text{ and not } \sigma^2.$$

Good estimates for σ^2 do exist, for example,

$$\frac{1}{2p}\sum_{j=1}^{p}(X_{j1}-X_{j2})^2 \xrightarrow[p\to\infty]{P} 2\sigma^2.$$

What is going wrong here?

This is not a *small p*, *large n* problem, but a *small n*, *large p* problem. i.e. a high dimensional problem, so needs care!

As $p \to \infty$, there are too many parameters to estimate and the likelihood function is unable to see where information lies, so tries to distribute it everywhere.

What is the way out? Go Bayesian!

There is a lot of information available on σ^2 (note $\sum_{j=1}^{p}(X_{j1} - X_{j2})^2 \sim 2\sigma^2\chi_p^2$) but very little on individual μ_j. However, if μ_j are 'similar', there is a lot of information on where they come from, because we get to see p samples, p large.

Suppose we are interested in μ_j. How can we use the above information? Model as follows:

$\bar{X}_j|\mu_j, \sigma^2 \sim N(\mu_j, \sigma^2/2)$, $j = 1, \ldots p$, independent observations. σ^2 may be assumed known, since a reliable estimate $\hat{\sigma^2} = \frac{1}{2p}\sum_{j=1}^{p}(X_{j1} - X_{j2})^2$ is available. Express the information that μ_j are 'similar' in the form:
$\mu_j, j = 1, \ldots p$ is a random sample (collection) from $N(\eta, \tau^2)$.
Where do we get the η and τ^2, the prior mean and prior variance?

Marginally (or in predictive sense) $\bar{X}_j, j = 1, \ldots p$ is a random sample from $N(\mu_0, \tau^2 + \sigma^2/2)$. Use this random sample.

Estimate η by $\hat{\eta} = \bar{\bar{X}} = \frac{1}{p}\sum \bar{X}_j$ and τ^2 by $\hat{\tau^2} = \left(\frac{1}{p-1}\sum_{j=1}^{p}(\bar{X}_j - \bar{\bar{X}})^2 - \sigma^2/2\right)^+$.

Now one could pretend that the prior for $(\mu_1, \ldots \mu_p)$ is $N(\hat{\eta}, \hat{\tau^2})$ and compute the Bayes estimates for μ_j:

$$E(\mu_j | \mathbf{X}_1, \ldots \mathbf{X}_p) = (1 - \hat{B})\bar{X}_j + \hat{B}\bar{\bar{X}},$$

where $\hat{B} = \frac{\sigma^2/2}{\sigma^2/2 + \hat{\tau^2}}$. If instead of 2 observations, each sample has n observations, replace 2 by n. This is called *Empirical Bayes* since the prior is estimated using data. There is also a fully Bayesian counter-part called Hierarchical Bayes.

Formal Methods for Model Selection

What is the best model for Gamma-ray burst afterglow?

Consider a simpler, abstract problem instead.
Suppose X having density $f(X|\theta)$ is observed, with θ being an unknown element of the parameter space Θ. We are interested in comparing two models M_0 and M_1:

$$\begin{aligned} M_0 &: X \text{ has density } f(x|\theta) \text{ where } \theta \in \Theta_0; \\ M_1 &: X \text{ has density } f(x|\theta) \text{ where } \theta \in \Theta_1. \end{aligned} \tag{11}$$

Simplify even further, and assume we want to test

$$M_0 : \theta = \theta_0 \text{ versus } M_1 : \theta \neq \theta_0, \tag{12}$$

Frequentist: A (classical) significance test is derived. It is based on a test statistic $T(X)$, large values of which are deemed to provide evidence against the null hypothesis, M_0. If data $X = x$ is observed, with corresponding $t = T(x)$, the P-value is

$$\alpha = P_{\theta_0}\left(T(X) \geq T(x)\right).$$

Example 6. Consider a random sample $X_1, \ldots, X_n$ from $N(\theta, \sigma^2)$, where σ^2 is known. Then $\bar{X}$ is sufficient for θ and it has the $N(\theta, \sigma^2/n)$ distribution. Noting that $T = T(\bar{X}) = |\sqrt{n}\left(\bar{X} - \theta_0\right)/\sigma|$ is a natural test statistic to test (12), one obtains the usual P-value as $\alpha = 2[1 - \Phi(t)]$, where $t = |\sqrt{n}\left(\bar{x} - \theta_0\right)/\sigma|$ and Φ is the standard normal cumulative distribution function.

What is a P-value and what does it say? P-value is the probability under a (simple) null hypothesis of obtaining a value of a test statistic that is at least as extreme as that observed in the sample data.

To compute a P-value we take the observed value of the test statistic to the reference distribution and check if it is likely or unlikely under M_0.

χ^2 Goodness-of-fit test

Example 7. Rutherford and Geiger (1910) gave the following observed numbers of intervals of 1/8 minute when 0, 1, ... α-particles are ejected by a specimen. Check if Poisson fits well.

Number	0	1	2	3	4	5	
Obs.	57	203	383	525	532	408	
Exp.	54	211	407	525	508	393	
Number	6	7	8	9	10	11	12 or more
Obs.	273	139	45	27	10	4	2
Exp.	254	140	68	29	11	4	1

Test statistic: $T = \sum_{j=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-2}$ approximately for large n,

where k is the number of cells, O_i is the observed and E_i is the expected count (estimated) for the ith cell.

Estimated Poisson intensity rate = (total number of particles ejected)/(total number of intervals) = 100097/2608 =3.87.
$k = 13$.
P-value = $P(T \geq 14.03) \approx 0.21$ (under χ^2_{11}).

Likelihood Ratio Criterion

Standard likelihood ratio criterion for comparing M_0 and M_1 is

$$\lambda_n = \frac{f(\mathbf{x}|\hat{\theta}_0)}{f(\mathbf{x}|\hat{\theta})} = \frac{\max_{\theta \in \Theta_0} f(\mathbf{x}|\theta)}{\max_{\theta \in \Theta_0 \cup \Theta_1} f(\mathbf{x}|\theta)}. \tag{13}$$

$0 < \lambda_n \leq 1$, and large values of λ_n provide evidence for M_0.
Reject M_0 for small values.
Use λ_n (or a function of λ_n) as a test statistic if its distribution under M_0 can be derived. Otherwise, use the large sample result:

$$-2\log(\lambda_n) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{p_1 - p_0},$$

under M_0 where p_0 and p_1 are the dimensions of Θ_0 and $\Theta_0 \cup \Theta_1$, respectively.

Bayesian Model Selection

How does the Bayesian approach work?

$X \sim f(x|\theta)$ and we want to test

$$M_0 : \theta \in \Theta_0 \quad \text{versus} \quad M_1 : \theta \in \Theta_1. \tag{14}$$

If Θ_0 and Θ_1 are of the same dimension (eg: $M_0 : \theta \leq 0$ and $M_1 : \theta > 0$), choose a prior density that assigns positive prior probability to Θ_0 and Θ_1. Then calculate the posterior probabilities $P\{\Theta_0|\mathbf{x}\}$, $P\{\Theta_1|\mathbf{x}\}$ as well as the posterior odds ratio, namely,

$$P\{\Theta_0|\boldsymbol{x}\}/P\{\Theta_1|\boldsymbol{x}\}.$$

Find a threshold like 1/9 or 1/19, etc. to decide what constitutes evidence against H_0.

Alternatively, let π_0 and $1 - \pi_0$ be the prior probabilities of Θ_0 and Θ_1. Let $g_i(\theta)$ be the prior p.d.f. of θ under Θ_i (or M_i), so that

$$\int_{\Theta_i} g_i(\theta) d\theta = 1.$$

The prior in the previous approach is nothing but

$$\pi(\theta) = \pi_0 g_0(\theta) I\{\theta \in \Theta_0\} + (1-\pi_0) g_1(\theta) I\{\theta \in \Theta_1\}. \qquad (15)$$

Need not require any longer that Θ_0 and Θ_1 are of the same dimension. Sharp null hypotheses are also covered. Proceed as before and report posterior probabilities or posterior odds. To compute these posterior quantities, note that the marginal density of X under the prior π can be expressed as

$$\begin{aligned} m_\pi(x) &= \int_\Theta f(x|\theta)\pi(\theta)\, d\theta \\ &= \pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta \qquad (16) \end{aligned}$$

and hence the posterior density of θ given the data $X = x$ as

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m_\pi(x)} = \begin{cases} \pi_0 f(x|\theta) g_0(\theta)/m_\pi(x) & \text{if } \theta \in \Theta_0; \\ (1-\pi_0) f(x|\theta) g_1(\theta)/m_\pi(x) & \text{if } \theta \in \Theta_1. \end{cases} \qquad (17)$$

It follows then that

$$
\begin{aligned}
P^{\pi}(M_0|x) &= P^{\pi}(\Theta_0|x) = \frac{\pi_0}{m_{\pi}(x)} \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta \\
&= \frac{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta}{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta} \quad \text{and} \\
P^{\pi}(M_1|x) &= P^{\pi}(\Theta_1|x) = \frac{(1-\pi_0)}{m_{\pi}(x)} \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta \\
&= \frac{(1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta}{\pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta)\, d\theta + (1-\pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta}.
\end{aligned}
$$

One may also report the *Bayes factor*, which does not depend on π_0. The Bayes factor of M_0 relative to M_1 is defined as

$$
BF_{01} = \frac{P(\Theta_0|\boldsymbol{x})}{P(\Theta_1|\boldsymbol{x})} \Big/ \frac{P(\Theta_0)}{P(\Theta_1)} = \frac{\int_{\Theta_0} f(\boldsymbol{x}|\theta) g_0(\theta)\, d\theta}{\int_{\Theta_1} f(\boldsymbol{x}|\theta) g_1(\theta)\, d\theta}. \qquad (18)
$$

Note:

- $BF_{10} = 1/BF_{01}$.
- Posterior odds ratio of M_0 relative to M_1:

$$\frac{P(\Theta_0|\boldsymbol{x})}{P(\Theta_1|\boldsymbol{x})} = \left(\frac{\pi_0}{1-\pi_0}\right) BF_{01}.$$

- Posterior odds ratio of M_0 relative to $M_1 = BF_{01}$ if $\pi_0 = \frac{1}{2}$.
- The smaller the value of BF_{01}, the stronger the evidence against M_0.

Testing as a model selection problem using Bayes factor illustrated below: Jeffreys test.

Jeffreys Test for Normal Mean; σ^2 Unknown

$X_1, X_2, \ldots, X_n$ a random sample from $N(\mu, \sigma^2)$. We want to test

$$M_0 : \mu = \mu_0 \text{ versus} M_1 : \mu \neq \mu_0$$

where μ_0 is some specified number.
Parameter σ^2 is common in the two models corresponding to M_0 and M_1 and μ occurs only in M_1. We take the prior $g_0(\sigma) = 1/\sigma$ for σ under M_0. Under M_1, we take the same prior for σ and add a conditional prior for μ given σ, namely

$$g_1(\mu|\sigma) = \frac{1}{\sigma} g_2(\frac{\mu}{\sigma}).$$

where $g_2(\cdot)$ is a p.d.f. Jeffreys suggested we should take g_2 to be Cauchy, so

$$g_0(\sigma) = \frac{1}{\sigma} \quad \text{under } M_0$$

$$g_1(\mu, \sigma) = \frac{1}{\sigma} g_1(\mu|\sigma) = \frac{1}{\sigma} \frac{1}{\sigma\pi(1 + \mu^2/\sigma^2)} \quad \text{under } M_1.$$

One may now find the Bayes factor BF_{01} using (18).

Example 8. Einstein's theory of gravitation predicts the amount of deflection of light deflected by gravitation. Eddington's expedition in 1919 (and other groups in 1922 and 1929) provided 4 observations: $x_1 = 1.98, x_2 = 1.61, x_3 = 1.18, x_4 = 2.24$ (all in seconds as measures of angular deflection). Suppose they are normally distributed around their predicted value μ. Then $X_1, \cdots, X_4$ are independent and identically distributed as $N(\mu, \sigma^2)$. Einstein's prediction is $\mu = 1.75$. Test $M_0 : \mu = 1.75$ versus $M_1 : \mu \neq 1.75$, where σ^2 is unknown.

Use the conventional priors of Jeffreys to calculate the Bayes factor. $BF_{01} = 2.98$.
The calculations with the given data lend some support to Einstein's prediction. However, the evidence in the data isn't very strong.

BIC

When we compare two models $M_0 : \boldsymbol{\theta} \in \Theta_0$ and $M_1 : \boldsymbol{\theta} \in \Theta_1$, what does the Bayes factor

$$BF_{01} = \frac{\int_{\Theta_0} f(\boldsymbol{x}|\theta) g_0(\theta)\, d\theta}{\int_{\Theta_1} f(\boldsymbol{x}|\theta) g_1(\theta)\, d\theta} = \frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})}$$

measure?

$m_0(\boldsymbol{x})$ measures how well M_0 fits the data $\boldsymbol{x}$ whereas $m_1(\boldsymbol{x})$ measures how well M_1 fits the same data, so BF_{01} is the relative strength of the two models in the predictive sense. This can be difficult to compute for complicated models, so any good approximation is welcome.

Approximate the marginal density $m(\boldsymbol{x})$ of $\boldsymbol{X}$ for large sample size n:

$$m(\boldsymbol{x}) = \int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta = ?$$

Laplace's Method

$$
\begin{aligned}
m(\boldsymbol{x}) &= \int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta = \int \pi(\theta) \prod_{i=1}^{n} f(x_i|\theta)\, d\theta \\
&= \int \pi(\theta) \exp\Big(\sum_{i=1}^{n} \log f(x_i|\theta)\Big)\, d\theta = \int \pi(\theta) \exp(nh(\theta))\, d\theta.
\end{aligned}
$$

where $h(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta)$.

Consider any integral of the form

$$
I = \int_{-\infty}^{\infty} q(\theta) e^{nh(\theta)}\, d\theta
$$

where q and h are smooth functions of θ with h having a unique maximum at $\hat{\theta}$.
If h has a unique sharp maximum at $\hat{\theta}$, then most contribution to the integral I comes from the integral over a small neighborhood $(\hat{\theta} - \delta, \hat{\theta} + \delta)$ of $\hat{\theta}$.

Study the behavior of I as $n \to \infty$. As $n \to \infty$, we have

$$I \sim I_1 = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} q(\theta) e^{nh(\theta)}\, d\theta.$$

Laplace's method involves Taylor series expansion of q and h about $\hat{\theta}$:

$$\begin{aligned} I &\sim \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \left[q(\hat{\theta}) + (\theta-\hat{\theta}) q'(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 q''(\hat{\theta}) + \cdots \right] \\ &\quad \times \exp\left[nh(\hat{\theta}) + nh'(\hat{\theta})(\theta-\hat{\theta}) + \frac{n}{2} h''(\hat{\theta})(\theta-\hat{\theta})^2 + \cdots \right] \\ &\sim e^{nh(\hat{\theta})} q(\hat{\theta}) \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \left[1 + (\theta-\hat{\theta}) q'(\hat{\theta})/q(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 q''(\hat{\theta})/q(\hat{\theta}) \right] \\ &\quad \times \exp\left[\frac{n}{2} h''(\hat{\theta})(\theta-\hat{\theta})^2 \right] d\theta. \end{aligned}$$

Assume $c = -h''(\hat{\theta}) > 0$ and use a change of variable $t = \sqrt{nc}(\theta - \hat{\theta})$:

$$
\begin{aligned}
I &\sim e^{nh(\hat{\theta})} q(\hat{\theta}) \frac{1}{\sqrt{nc}} \\
&\quad \times \int_{-\delta\sqrt{nc}}^{\delta\sqrt{nc}} \left[1 + \frac{t}{\sqrt{nc}} q'(\hat{\theta})/q(\hat{\theta}) + \frac{t^2}{2nc} q''(\hat{\theta})/q(\hat{\theta}) \right] e^{-t^2/2}\, dt \\
&\sim e^{nh(\hat{\theta})} \frac{\sqrt{2\pi}}{\sqrt{nc}} q(\hat{\theta}) \left[1 + \frac{q''(\hat{\theta})}{2ncq(\hat{\theta})} \right] \\
&= e^{nh(\hat{\theta})} \frac{\sqrt{2\pi}}{\sqrt{nc}} q(\hat{\theta}) \left[1 + O(n^{-1}) \right]. \qquad (19)
\end{aligned}
$$

Apply (19) to $m(\boldsymbol{x}) = \int \pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta = \int \pi(\theta) \exp(nh(\theta))\, d\theta$, with $q = \pi$ and ignore terms that stay bounded.

$\log(m(\boldsymbol{x}) \approx nh(\hat{\theta}) - \frac{1}{2} \log n = \log(f(\boldsymbol{x}|\hat{\theta})) - \frac{1}{2} \log n.$

Schwarz (1978) proposed a criterion, known as the BIC, based on (19) ignoring the terms that stay bounded as the sample size $n \to \infty$ (and general dimension p for θ):

$$BIC = \log f(\boldsymbol{x}|\hat{\theta}) - (p/2)\log n$$

This serves as an approximation to the logarithm of the integrated likelihood of the model and is free from the choice of prior.

$2\log B_{01}$ is a commonly used evidential measure to compare the support provided by the data $\mathbf{x}$ for M_0 relative to M_1. Under the above approximation we have,

$$2\log(B_{01}) \approx 2\log\left(\frac{f(\mathbf{x}|\widehat{\theta}_0)}{f(\mathbf{x}|\widehat{\theta}_1)}\right) - (p_0 - p_1)\log n. \qquad (20)$$

This is the approximate Bayes factor based on the Bayesian information criterion (BIC) due to Schwarz (1978). The term $(p_0 - p_1)\log n$ can be considered a penalty for using a more complex model.

AIC

Recall the likelihood ratio criterion: $\lambda_n = \frac{f(\mathbf{x}|\hat{\theta}_0)}{f(\mathbf{x}|\hat{\theta})}$

$$P(M_0 \text{ is rejected}|M_0) = P(\lambda_n < c) \approx P(\chi^2_{p_1-p_0} > -2\log(c)) > 0,$$

so, from a frequentist point of view, a criterion based solely on the likelihood ratio does not converge to a sure answer under M_0.

Akaike (1983) suggested a penalized likelihood criterion:

$$2\log\left(\frac{f(\mathbf{x}|\widehat{\theta}_0)}{f(\mathbf{x}|\widehat{\theta}_1)}\right) - 2(p_0 - p_1) \tag{21}$$

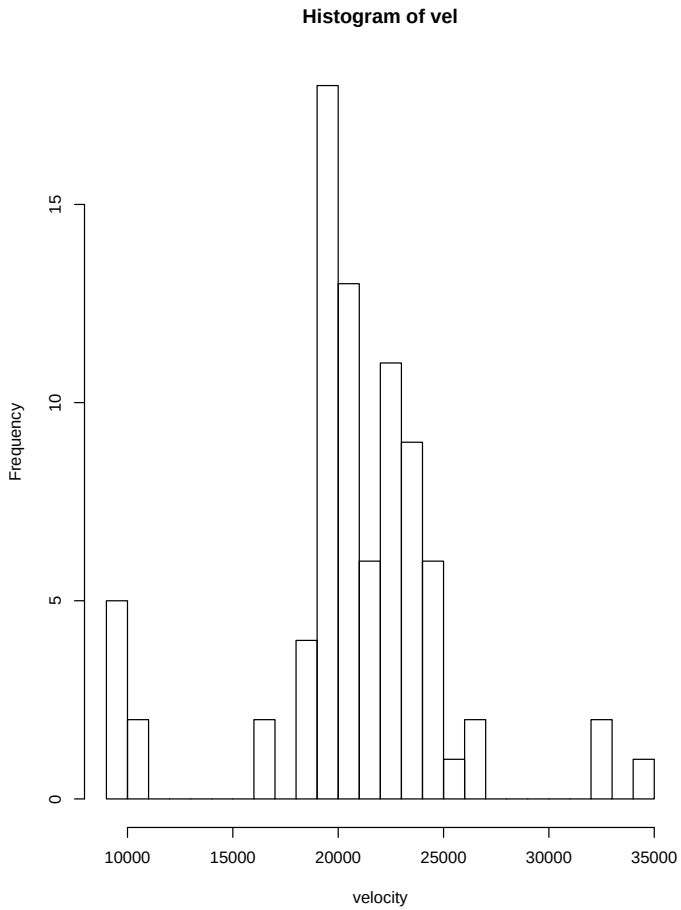
which is based on the Akaike information criterion (AIC), namely,

$$AIC = 2\log f(\mathbf{x}|\widehat{\theta}) - 2p$$

for a model $f(\mathbf{x}|\theta)$. The penalty for using a complex model is not as drastic as that in BIC.

Model Selection or Model Averaging?

Example 9. Velocities (km/second) of 82 galaxies in six well-separated conic sections of the Corona Borealis region. How many clusters?



Consider mixture of normals:

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \left(\sum_{j=1}^{k} p_j \phi(x_i|\mu_j, \sigma_j^2) \right), \end{aligned}$$

where k is the number of mixture components, p_j is the weight given to the jth component, $N(\mu_j, \sigma_j^2)$.

Models to consider:

$$M_k : X \text{ has density } \sum_{j=1}^{k} p_j \phi(x_i|\mu_j, \sigma_j^2), k = 1, 2 \ldots$$

i.e., M_k is a k component normal mixture.

Bayesian model selection procedure computes $m(\boldsymbol{x}|M_k) = \int \pi(\boldsymbol{\theta}_k) f(\boldsymbol{x}|\boldsymbol{\theta}_k)\, d\boldsymbol{\theta}_k$, for each k of interest and picks the one which gives the largest value.

Example 9 contd. Chib (1995), JASA:

k	σ_j^2	$\log(m(\boldsymbol{x}\|M_k))$
2	$\sigma_j^2 = \sigma^2$	-240.464
3	$\sigma_j^2 = \sigma^2$	-228.620
3	σ_j^2 unrestricted	-224.138

3 component normal mixture model with unequal variances seems best.

- From the Bayesian point of view, a natural approach to model uncertainty is to include all models, M_k, under consideration for future decisions.
- i.e., Bypass the model-choice step entirely.
- Unsuitable for scientific inference where selection of a model is a must.
- Suitable for prediction purposes, since underestimation of uncertainty resulting from choosing model $M_{\hat{k}}$ is eliminated.

Predictive density $m(y|\boldsymbol{x})$ given the sample $\boldsymbol{x} = (x_1, \ldots x_n)$ is obtained by averaging over all models:

$$\begin{aligned} m(y|\boldsymbol{x}) &= \int_\Theta f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} \\ &= \sum_k \int_{\Theta_k} f_k(y|\boldsymbol{\theta}_k)\pi(k, \boldsymbol{\theta}_k|\boldsymbol{x})\, d\boldsymbol{\theta}_k \\ &= \sum_k P(M_k|\boldsymbol{x}) \int_{\Theta_k} f_k(y|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\boldsymbol{x})\, d\boldsymbol{\theta}_k, \end{aligned}$$

where $\Theta = \cup_k \Theta_k$ and

$$f(y|\boldsymbol{\theta}) = \sum_k p_k f_k(y|\boldsymbol{\theta}_k).$$

References

1. Tom Loredo's site: `http://www.astro.cornell.edu/staff/loredo/bayes/`
2. *An Introduction to Bayesian Analysis: Theory and Methods* by J.K. Ghosh, Mohan Delampady and T. Samanta, Springer, 2006
3. *Probability Theory: The Logic of Science* by E.T. Jaynes, Cambridge University Press, 2003
4. Bayesian Logical Data Analysis for Physical Sciences by P.C. Gregory, Cambridge University Press, 2005
5. *Bayesian Reasoning in High-Energy Physics: Principles and Applications* by G. D'Agostini, CERN, 1999