BAYESIAN INFERENCE

BY

TANUJIT CHAKRABORTY

Indian Statistical Institute

Mail : tanujitisi@gmail.com

BAYESIAN INFERENCE

$f(x|\theta)$ x: Effect θ: Cause : Probability thinking

$l(\theta|x)$ Inversion process.

Bayesian thinking $\Longleftrightarrow$ Inverse probability thinking.

Historical Background:- Probability has been a subject of study for a long time. Statistics is relatively a young field. Linear regression due to Galton appeared in the late 1800's and Karl Pearson's goodness of fit measure and correlation appeared in 19th century.

The field of Statistics blossomed in the 1920's and 1930's with works of Fisher, Neyman, Pearson. A flurry of research was prompted by World War 2, which generated a wide variety of difficult applied problems and the first substantive government funding came in US and UK.

By contrast, Bayesian methods are much older, dating to the original paper of 'THOMAS BAYES' in the 1760s. The area generated some interest in the works of Laplace, Gauss and others in the 19th century. The Bayesian approach was ignored and opposed by the statisticians of the early 20th century. Fortunately, during this period, several prominent non-statisticians, most notably Harold Jeffreys (a Physicist) and Arthur Bowley (an Econometrician) continued to lobby on behalf of Bayesian ideas (which they referred to as "Inverse Probability").

Beginning 1950's, statisticians such as L.J.Savage, Bruno de Finetti, Dennis Lindley and many others come in and they pointed out several deficiencies of the classical approach. The biggest impetus to Bayesian statistics came in 1990's after many computational algorithms and modern computers started to surface.

Motivating Examples:- "Acceptance of paper in Anals of Statistics".

What is the probability of accepting your paper with a prior information that your 5 papers submitted to Anals got accepted?

Some Paradoxes:—

- **Example 1:** $X_1, X_2 \overset{iid}{\sim} U\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, $\theta \in \mathbb{R}$.

A 95% CI for θ is given by $[\bar{X} - c, \bar{X} + c]$ by choosing c properly.

Now let us suppose we are given $X_{(2)} = 2$ and $X_{(1)} = 1$.

Then θ is obviously $\frac{3}{2} = \bar{X}$.

But a frequentist, if he sticks to the long term frequency interpretation of CI. will not be able to say if the given interval contains θ or not surely !!

- **Example 2:** (Cox 1958)

To estimate μ in $N(\mu, \sigma^2)$, toss a fair coin. Draw a random sample of size $n = 2$ if it falls heads and draw a sample of size $n = 1000$ if it falls tails. An unbiased estimate of μ is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ with variance $\frac{1}{2}\left(\frac{\sigma^2}{2} + \frac{\sigma^2}{1000}\right) \simeq \frac{\sigma^2}{4}$.

Will you believe that the variance is like $\frac{\sigma^2}{4\bar{X}}$ if you actually draw a sample of size 1000 and compute based on that sample ?

- **Example 3:** (Lindley and Philips (1976)):

Suppose in 12 independent tosses of a coin you observe 9 heads and 3 tails and the last toss producing a tail.

To test $H_0 : \theta = \frac{1}{2}$ Vs. $H_1 : \theta > \frac{1}{2}$ (θ: Prob. of success)

We apply ~~two~~ choices of possible sampling distributions.

(a) **Binomial:** Here $n = 12$ tosses are fixed before hand and the random quantity X is the # of heads obtained in n tosses.

Then $X \sim \text{Bin}(12, \theta)$.

Likelihood (for the given data) is given by $\binom{12}{9}\theta^9(1-\theta)^3$.

(b) **Negative Binomial:** Here one has to assume that data collection continued until this tail appeared. The random quantity X is the number of heads required to complete the experiment so that $X \sim \text{Neg. Bin}(r = 3, \theta)$.

Likelihood is given by $\binom{11}{9}\theta^9(1-\theta)^3$.

Consider the Rejection Region: Reject H_0 if $X \geq c$.

p-value :-

Binomial case: $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = 0.075$

Neg-Bino. case: $\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{\infty} \binom{12+j}{j} \theta^j (1-\theta)^3 = 0.0325.$

Using the "usual" Type I error level $\alpha = 0.05$, two models lead to two decisions. You should note that the likelihoods are proportional here.

One of the main thing what went 'wrong' is that we allowed unobserved outcomes to affect the rejection decision. That is, probability of X values may be greater than 9 (the value actually observed) was used as evidence against H_0 in each case, even though these values did not occur.

As mentioned before, statistical inference is based on a probabilistic modelling of the observed phenomenon. There are two approaches:

— A first approach assumes that statistical inference must incorporate as much as possible of the complexity of the phenomenon, and thus aims at estimating the distributions underlying the phenomenon under minimal assumptions, generally using functional estimation (density, regression, etc). This is non-parametric approach.

— conversely, the parametric approach represents the distribution through a density function $f(x|\theta)$, where only the parameter θ (of finite dimension) is unknown. This is the parametric approach.

- Reference Books:- 1. An Introduction to Bayesian Analysis by JK Ghosh, Mohan Delampady and Tapas Samanta
2. Bayesian Data Analysis 3 by Rubin, Stern, Gelman, Dunson
3. Monte Carlo Statistical Methods by Robert Casella.
4. Collected works of D. Basu by JK Ghosh, revised version by Anirban Dasgupta.

Classical Inference Model:-

Let us consider for example, the problem of estimation of a real parameter.

θ is unknown in nature.

Data: $\underset{\sim}{X} \sim P_\theta$, $\theta \in (H)$

$T(\underset{\sim}{X})$: estimator

For a good estimator, $T(\underset{\sim}{X})$, $|T(\underset{\sim}{X}) - \theta|$ should be small or equivalently $(T(\underset{\sim}{X}) - \theta)^2$ should be small.

Classical statistics averages this quantity over the whole sample space $\mathcal{X}$ (set of all possible values of $\underset{\sim}{X}$)

$$R(T, \theta) = E_\theta \left(T(\underset{\sim}{X}) - \theta\right)^2$$

A Bayesian won't do this. He/she will consider a data dependent measure of performance of an estimator — something that depends on the particular data in hand, the data he has observed. Once he has a particular data that he has observed, he is not going to average over all other possible values of the data that he have not observed.

Classical Statistics:- An estimator T_0 is said to be the best if it minimizes $R(T, \theta)$ w.r.t. choices of the estimator T, Here θ is unknown. So, this is a very difficult problem. Indeed, there does not exist a "best" estimator.

There are two ways:-

(i) Imposing restrictions to the class of estimators (such as unbiasedness and invariance) and finding the best estimator in the restricted class.

(ii) Considering all the estimators but evaluating them in a different manner. (leading to the minimax estimators)

e.g. $\sup_\theta R(T, \theta)$

Example:- Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$; $\theta \in \mathbb{R}$

UMVUE is $\bar{X}$. Suppose we know $-1 \le \theta \le 1$. Then UMVUE is a bad estimator. There are natural common sense estimators of θ better than $\bar{X}$.

The Bayesian Model:- The purpose of statistical analysis is fundamentally an inversion purpose, since it aims at retrieving the causes — reduced to the parameter of the probabilistic generating mechanism from the effects — summarized by the observations.

In other words, when observing a random phenomenon directed by a parameter θ, statistical methods allow to deduce from these observations an inference (i.e., a summary of a characterization) about θ.

A general description of inversion of probabilities is given by Bayes formula.

If A and E are events such that $P(E) \neq 0$, $P(A|E)$ and $P(E|A)$ are related by

$$P(A|E) = \frac{P(E|A)\,P(A)}{P(E|A)\,P(A) + P(E|A^c)\,P(A^c)}.$$

This result describes the updating of likelihood of A from $P(A)$ to $P(A|E)$ once E has been observed. Thomas Bayes actually proved a continuous version of the result, namely, given x and y with conditional distribution $f(x|y)$ and marginal $g(y)$, the conditional distribution of y given x is

$$g(y|x) = \frac{f(x|y)\,g(y)}{\int f(x|y)\,g(y)\,dy}.$$

While this inversion is quite natural from a probabilistic point of view; Bayes and Laplace went further and considered that the uncertainty of the parameter θ of a model could be modelled through a probabilistic distribution π on Θ (parameter space) called the prior distribution.

The inference then based on the distribution of θ conditional on x, say, $\pi(\theta|x)$, called the posterior distribution, given by

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

The distribution $\pi(\theta)$ is called a prior distribution on a prior because it quantifies one's uncertainty about θ, prior to observing data. The prior may represent the Bayesian's subjective belief/knowledge, in which case it would be a subjective prior.

Alternatively, it could be a conventional prior supposed to represent small or no information.

Recall $$\pi(\theta|x) = \frac{\pi(\theta)\,f(x|\theta)}{\int \pi(\theta)\,f(x|\theta)\,d\theta} \quad \text{--------(*)}$$

(H)

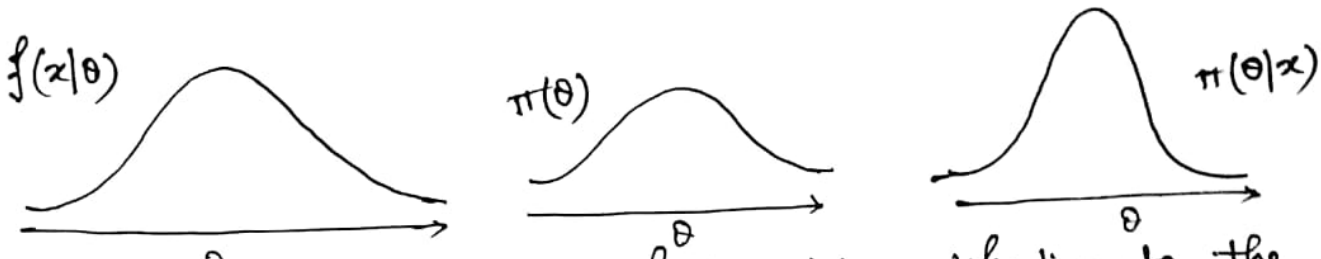
i.e., $$\pi(\theta|x) \propto \pi(\theta)\,f(x|\theta).$$

The numerator in (*) is the joint density of θ and x.

The denominator $m(x) \overset{\text{def}}{=} \int \pi(\theta)\,f(x|\theta)\,d\theta$ is the marginal density of x.

The posterior distribution combines the prior beliefs about θ with the information about θ contained in the data x, to give a composite picture of the final beliefs about θ.

Just as prior distribution reflects beliefs about θ prior to analysing data, $\pi(\theta|x)$ reflects the updated beliefs about θ after (posterior to) observing data x.



In summary, by imposing a fundamental modification to the perception of random phenomena, in the sense that parameters diverting random phenomena can also be perceived as "random variables". Bayes and Laplace created "modern" statistical analysis. The use of prior is the best way to summarize the available information (or even lack of information) about θ as well as residual uncertainty, thus allowing for incorporation of this imperfect information in the decision process.

Remark:- The posterior distribution contains all the current information about θ. Ideally, one might report the entire posterior distribution $\pi(\theta|x)$ or one could report descriptive summary measures associated with the posterior distribution.

Example:- For a real valued parameter θ, one could report the parameter mean $E(\theta|x) = \int \theta \pi(\theta|x)\, d\theta$ and the

(H)

posterior variance $Var(\theta|x) = E\left[\left(\theta - E(\theta|x)\right)^2 \middle| x\right]$

$$= \int \left(\theta - E(\theta|x)\right)^2 \pi(\theta|x)\, d\theta.$$

Finally one could use $\pi(\theta|x)$ to answer question related to more structured problem like estimation or testing.

Point Estimation:- Commonly used summaries of location — like mean, median or mode of $\pi(\theta|x)$ may be used as point estimates of θ. Together with a point estimate $\hat{\theta}$ of θ, one also reports a measure of extent of uncertainty associated with the estimate.

Example: Together with the posterior mean one also reports posterior variance.

Optimum estimates can be obtained for given loss function. Let, $L(\theta, a)$ be the loss in estimating θ by a. Bayes estimate of θ is obtained by minimizing the average loss

$$E^{\pi(\cdot|x)} L(\theta, a) = \int L(\theta, a)\pi(\theta|x)\,d\theta \quad \text{w.r.t. } a.$$

The minimizer is called the Bayes estimate of θ.

For $L(\theta, a) = (\theta - a)^2$, then Bayes estimate is posterior mean,

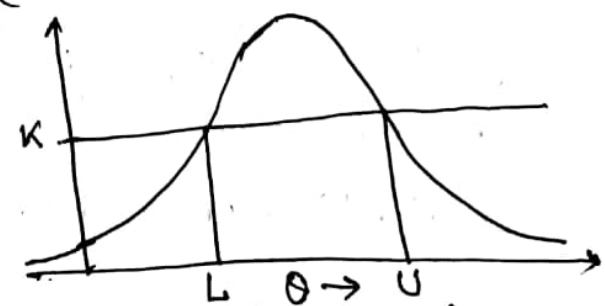
For $L(\theta, a) = |\theta - a|$, then Bayes estimate is posterior median.

If one reports posterior median, it is convenient to report a couple of other posterior quantities to give an idea of the posterior variability of θ. One may report the first and third posterior quantiles.

Confidence Set:- The Bayesian analysis of classical confidence set is called a credible set.

Definition:- For $0 < \alpha < 1$, a $100(1-\alpha)\%$ credible set for θ is a subset C of the parameter space (H) w.r.t.

$$P(\theta \in C|\underset{\sim}{x}) = \int_{\theta \in C} \pi(\theta|x)\,d\theta \geq 1 - \alpha.$$



HPD : $\{\theta : \pi(\theta|x) \geq K\}$

Where as the classical confidence statement don't apply to whether for a given set for a given x covers the "true" θ, this is not the case with credible set does answer a layman's question whether the given set covers the "true" θ with probability $1-\alpha$. This is because, in the Bayesian approach, "true" θ is a random variable, with a data dependent probability distribution (after data has been observed), namely, the posterior distn..

In choosing a credible set for θ, it is usually desirable to minimize its size. To this, one should include in the set only those points with the largest posterior density, i.e., the "most likely" values of θ.

Definition:- The $100(1-\alpha)\%$ Highest Posterior Density (HPD) credible set for θ is the subset C of Θ of the form

$$C = \{\theta \in \Theta : \pi(\theta|x) \geq K(\alpha)\}, \text{ where}$$

$K(\alpha)$ is the largest constant respect to $P(\theta \in C \mid x) \geq 1-\alpha$.

Example of credible Sets:- $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$; σ^2 known.

Let $\theta \sim N(\mu, \tau^2)$ as before (μ, τ^2) is known, then highest posterior density (HPD) credible set for θ is an equal tailed interval C centred at the posterior mean

$$C = E(\theta|\underset{\sim}{x}) \pm z_{\alpha/2}\sqrt{\text{var}(\theta|\underset{\sim}{x})}$$

Example:- $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Cauchy}(\theta)$; $\theta > 0$

$$f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)}; \quad -\infty < x < \infty, \text{ where } \theta \text{ is the location parameter.}$$

A reasonable objective prior one can easily find a $100(1-\alpha)\%$ HPD credible set using computer numerically based on posterior densities

$$\pi(\theta|\underset{\sim}{x}) \propto \prod_{i=1}^{n}\left[1+(\theta-x_i)^2\right]^{-1}; \quad \theta > 0.$$

In contrast, it is not at all clear how to construct a classical confidence set for this problem for fixed sample size situation. For very large sample sizes, we may use the asymptotic normality of sample median to produce an asymptotically correct frequentist CI for θ.

Hypothesis Testing:- It is possible to derive posterior probability of a hypothesis H_0, i.e., $P^{\pi}(H_0 | \underset{\sim}{x})$. Suppose we want to test $H_0: \theta \leq \theta_0$ Vs. $H_1: \theta > \theta_0$. One calculates

$$P(H_0 | \underset{\sim}{x}) = P(\theta \leq \theta_0 | \underset{\sim}{x}) = \int_{\theta \leq \theta_0} \pi(\theta | \underset{\sim}{x}) d\theta.$$

Exercise:- Consider the problem of estimation of a real parameter θ with the loss function:

$$L(\theta, a) = K_0(\theta - a) \text{ if } \theta - a \geq 0$$
$$= K_1(a - \theta) \text{ if } \theta - a < 0$$

Show that the Bayes estimate is given by the quantity of order $\frac{K_0}{K_0 + K_1}$ of the posterior distribution (assume, for simplicity uniqueness of the quantile).

Example 1: Normal Means:-

Hypothesis Testing:- Typically for tests like $H_0: \theta \in \Theta_0$ Vs. $H_1: \theta \in \Theta - \Theta_0$, one derives the testing rule using the posterior probability $P^{\pi}(H_0 | \underset{\sim}{x})$ on H_0 being true.

Example: 1. Normal Means:-

Suppose given θ; $x_1, x_2, \ldots, x_n$ are iid $N(\theta, \sigma^2)$ where σ^2 is the known parameter and the parameter of interest is θ; say, $\theta \in \mathbb{R}$.

A convenient and desirable prior distn. for θ is a normal distn. with suitable prior mean and variance, say, $\theta \sim \pi(\theta) = N(\mu, \tau^2)$ where, μ and τ^2 are both known. The prior variance τ^2 is a measure of strength of our belief in the prior mean in the sense that larger the τ^2 the less sure. We are about our prior guess. The likelihood is —

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

Prior density, $\pi(\theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left[-\frac{1}{2\tau^2}(\theta-\mu)^2\right]$

Posterior density, $\pi(\theta|\underset{\sim}{x}) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta}$

$\propto f(x|\theta)\pi(\theta) = \exp\left(-\frac{M}{2}\right)$, where,

$$M = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 + \frac{1}{\tau^2}(\theta-\mu)^2$$

$$= \frac{1}{\sigma^2}\left(\sum x_i^2 - 2\theta\sum x_i + n\theta^2\right) + \frac{1}{\tau^2}\left(\theta^2 - 2\theta\mu + \mu^2\right)$$

$$= \theta^2\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu}{\tau^2}+\frac{\sum x_i}{\sigma^2}\right) + \frac{\mu^2}{\tau^2} + \frac{\sum x_i^2}{\sigma^2}.$$

$$= \theta^2 a - 2b\theta + c$$

$$= a\left(\theta - \frac{b}{a}\right)^2 + c - \frac{b^2}{a}.$$

Thus, $\pi(\theta|\underset{\sim}{x}) \propto \exp\left[-\frac{1}{2}a\left(\theta-\frac{b}{a}\right)^2\right]$

$$\therefore \theta|\underset{\sim}{x} \sim N\left(\frac{b}{a}, \frac{1}{a}\right) = N\left(\frac{\frac{\sum x_i^2}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right)$$

$$E(\theta|\underset{\sim}{x}) = \frac{\frac{\mu}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} \quad \text{and} \quad \text{Var}^{\pi}(\theta|\underset{\sim}{x}) = \frac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}.$$

(posterior mean) (posterior variance)

If precision is measured by reciprocal of variance, posterior precision is $\frac{1}{\tau^2}+\frac{n}{\sigma^2}$. In the layout of the data, the precision in the prior knowledge has increased and is reflected in the increased value of the posterior precision.

Note that, the posterior mean θ is an weighted average of the prior estimate μ and the classical estimate $\bar{x}$ with weights being proportional to the corresponding precisions

$$E(\theta|\underset{\sim}{x}) = \left(\frac{1/\tau^2}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right)\mu + \left(\frac{n/\sigma^2}{1/\tau^2+\frac{n}{\sigma^2}}\right)\bar{x}.$$

It may be considered as a modification of the classical estimate $\bar{x}$ in the light of prior information about θ or as a modification of the prior estimate μ of θ in the light of the data and note that as $\tau \to \infty$; $E(\theta|\underset{\sim}{x}) \to \bar{x}$ as $n \to \infty$.
Thus if the prior information is small or there are lots of data points the posterior mean is close to the classical estimate $\bar{x}$.
In both cases $\pi(\theta|\underset{\sim}{x}) \approx N\left(\bar{x}, \sigma^2/n\right)$; as $n \to \infty$, $V(\theta|\underset{\sim}{x}) \to 0$.
Thus as $n \to \infty$ the posterior distn becomes more and more concentrated around $\bar{x}$.

- The posterior depends both on prior and x.
- As amount of data increases the influence of data leads to wash away the prior.

Example : 2. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p)$; let p have a prior distn $\pi(p)$.
We will consider a family of priors for p which simplifies the calculation of posterior and then consider some commonly used priors from this family.
Let $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$; $\text{Beta}(\alpha, \beta)$; $\alpha > 0$, $\beta > 0$.

$$\text{Prior mean} = \frac{\alpha}{\alpha+\beta} \text{ ; Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The posterior, $\pi(p|\underset{\sim}{x}) \propto p^{\alpha+\gamma-1}(1-p)^{\beta+(n-\gamma)-1}$, where $\gamma = \sum_{i=1}^{n} X_i$.
The posterior distn of p is $\text{Beta}(\alpha+\sum X_i, \beta+n-\sum X_i)$.

$$E(p|\underset{\sim}{x}) = \frac{\alpha+\sum X_i}{\alpha+\beta+n}$$

$$\text{Var}(p|\underset{\sim}{x}) = \frac{(\alpha+\sum X_i)(\beta+n-\sum X_i)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}.$$

If n is large then $E(p|\underset{\sim}{x}) \approx \hat{p} = \frac{\sum X_i}{n}$ is the classical estimate and $\text{Var}(p|\underset{\sim}{x})$ is also quite small. The posterior distn is concentrated around $\hat{p} = \sum X_i/n$.
We can interpret this as an illustration of a fact mentioned above: when we have lots of data, the data tend to wash away the influence of prior. The posterior mean can be written as a weighted average of the prior mean (prior estimate) and the classical estimate (MLE) $\hat{p} = \frac{\sum X_i}{n}$.

Note that the posterior mean $E\left(p \mid \underset{\sim}{X}\right)$ can be written as

$$\frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right)+\frac{n}{\alpha+\beta+n}\left(\frac{\sum X_i}{n}\right).$$

- As in example 1, we see that the prior mean and the classical estimate are being combined in a convex way.
- Taking $\alpha=1=\beta$, for the uniform distn., the posterior mean is $\frac{\sum X_i+1}{n+2}$. The prior was used by Bayes and Laplace. (Uniform prior)
- If $\alpha=\beta=1/2$, we have what is called the Jeffrys's prior. Jeffrys prior $\pi(\theta) \propto \sqrt{\det\left(I(\theta)\right)}$; where $I(\theta)$ is the fisher information based on single observation.
- The posterior mean corresponding to Jeffry's prior is given by $\frac{\sum X_i + 1/2}{n+1}$. Jeffry's prior is a very widely used and popular prior in the case of one dimensional parameter.
- When $\alpha=\beta=0$, the posterior mean is equal to MLE $= \frac{\sum X_i}{n}$.

Remark:- If we had wished to test $H_0: \mu=\mu_0$ vs. $H_1: \mu \neq \mu_0$ or $H_1: \mu > \mu_0$. We have to choose the prior in a careful way since any continuous prior on μ would give zero probability to H_0. In these problems the bayesian answers tend to differ substancially from classical answers.

$X_1, \ldots, X_n \overset{iid}{\sim} C(\theta, 1)$

$X_i - \theta \overset{iid}{\sim} C(0,1)$

$\bar{X} - \theta \overset{iid}{\sim} C(0,1)$

Improper Priors:- Consider the "normal mean" example:

$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, σ^2 is known, $\theta \sim N(\mu, \tau^2)$.

As τ^2 becomes larger, the prior becomes more and more diffuse/flat. If τ^2 is very large ($\tau^2 \to \infty$),

$$\pi(\theta \mid \underset{\sim}{x}) \simeq N(\bar{x}, \sigma^2/n)$$ density for θ.

Putting this in another way, if we let $\tau^2 \to \infty$, then in the limiting case it would give an <u>improper</u> uniform prior over the whole real line

$$\pi(\theta) = c, \quad -\infty < \theta < \infty$$ (c is an arbitrary constant)

which, by the Bayes formula, will give a $N(\bar{x}, \frac{\sigma^2}{n})$ posterior for θ.

Note that $\int \pi(\theta)\, d\theta = \infty$. Such a prior distn is called <u>improper</u>. We call a prior density $\pi(\theta)$ proper if it integrates to 1.

If $\int \pi(\theta)\, d\theta$ is finite, it is called an unnormalised density and can be normalized — multiplied by a constant — to integrate to 1.

In the above example, despite the impropriety of the prior distn., the posterior distn. is proper given just one observation.

With an improper prior, we can proceed with the Bayes formula and define an unnormalized posterior density by

$$\pi(\theta \mid \underset{\sim}{x}) \propto f(\underset{\sim}{x} \mid \theta)\, \pi(\theta).$$

If $m(\underset{\sim}{x}) = \int f(\underset{\sim}{x} \mid \theta)\, \pi(\theta)\, d\theta < \infty$ this leads to a proper posterior density. Bayesian methods apply as long as the posterior distribution is proper.

Noninformative Priors (Objective Priors):-

If no prior information is available, or, it is very difficult to quantify the available information into a prior distribution, one can still use Bayesian methods using what are called noninformative ("objective") priors.

Standard noninformative priors are:

Uniform prior : $\pi(\theta) = \text{constant}$

Jeffreys prior : $\pi(\theta) = \sqrt{\det(I(\theta))}$, where $I(\theta)$ is fisher Information matrix.

Reference priors (Bernardo, Berger)

Several methods for construction of noninformative priors are available.

Standard noninformative priors for location parameter (μ) and scale parameter (σ) are: $\pi(\mu) = c$, $-\infty < \mu < \infty$ and $\pi(\sigma) \propto \frac{1}{\sigma}$, $0 < \sigma < \infty$.

$$f(x, \theta) = f(x - \theta), \quad f(x, \sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right), \quad \sigma > 0.$$

Note: Objective priors are usually improper. To be usable they must have proper posteriors. The uniform, Jeffrey's prior and reference priors are examples of objective priors. All of them produce a posterior mean that is close to the MLE mean for small n.

Conjugate Priors:- In the Normal and Bernoulli examples, we saw that by taking normal and Beta as priors respectively the posterior remains within the normal and beta families. The property that posterior distn is of the same parametric form as the prior distn is called conjugacy. The corresponding priors are called conjugate priors. These priors also have great advantage in terms of computational case and can also be interpreted as "additional data".

Example: ① $X_1, \ldots, X_n \overset{iid}{\sim}$ Poisson(θ); $\theta > 0$.

$$\pi(\theta) = \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma\alpha} \; ; \; \theta > 0 \qquad [\text{Gamma}(\alpha, \beta)]$$

Posterior density $\propto \theta^{\sum x_i + \alpha - 1} e^{-(\beta+n)\theta}$

i.e., Gamma $(\alpha + \sum x_i, \beta + n)$

$$E(\theta|\underset{\sim}{x}) = \frac{\alpha + \sum x_i}{\beta + n} \; ; \; \text{Var}(\theta|\underset{\sim}{x}) = \frac{\alpha + \sum x_i}{(\beta+n)^2}.$$

② $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$; μ known, σ^2 unknown.

Likelihood is $f(x|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left[-\frac{nv}{2\sigma^2}\right]$;

where $v = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$

The conjugate prior density is the inverse gamma.

$$\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

Posterior: $\pi(\sigma^2|\underset{\sim}{x}) \propto (\sigma^2)^{-(\alpha+1+n/2)} \exp\left[-\frac{1}{\sigma^2}\left(\beta + \frac{nv}{2}\right)\right]$

$$L(\sigma^2, a) = \frac{(a-\sigma^2)^2}{\sigma^4} = \left(\frac{a}{\sigma^2} - 1\right)^2.$$

To minimize w.r.t. a,

$$\int L(\sigma^2, a)\pi(\sigma^2|\underset{\sim}{x})\, d\sigma^2$$

$$= \int (a-\sigma^2)^2 (\sigma^2)^{-2} \times \{\text{density of IG}\left(\alpha + \tfrac{n}{2}, \beta + \tfrac{nv}{2}\right) \text{ for } \sigma^2\} d\sigma^2$$

$$= \text{constant} \times \int (a-\sigma^2)^2 \times \{\text{density of IG}\left(\alpha + \tfrac{n}{2} + 2, \beta + \tfrac{nv}{2}\right) \text{ for } \sigma^2\} d\sigma^2$$

$E\{L(\sigma^2, a) | X\}$ is minimized at

$$a = \text{Mean of } IG\left(\alpha + 2 + \frac{n}{2}, \beta + \frac{n\upsilon}{2}\right)$$

$$= \frac{\beta + \frac{n\upsilon}{2}}{\alpha + 1 + \frac{n}{2}} = \text{mode of posterior dist}^n.$$

[If $\theta \sim IG(\alpha, \beta)$; $E(\theta) = \frac{\beta}{\alpha - 1}$; $\alpha > 1$

$$V(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} ; \quad \alpha > 2$$

$$\text{Mode}(\theta) = \frac{\beta}{\alpha + 1} .]$$

(3) Exponential Family:- $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$

$f(x|\theta) = \exp\{\theta t(x) + c(\theta)\} h(x)$; $\theta \in (a, b)$; a, b could be $\pm\infty$.

Assume all regularity conditions and smoothness of $c(\theta)$; consider the problem of estimating $\mu = E_\theta(t(X))$.

Note:- $\mu = \mu(\theta) = -c'(\theta)$

$$E_\theta\left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] = 0 = E_\theta\left[c'(\theta) + t(X)\right]$$

$$\Rightarrow E_\theta(t(X)) = -c'(\theta).$$

Now, $0 < I(\theta) = \text{Var}\left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right]$

$$= -E_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right]$$

$$= -c''(\theta)$$

i.e., $\frac{d\mu}{d\theta} = -c''(\theta) > 0$ implying μ is a strictly increasing function of θ.

MLE of μ:

$L = f(x_1, \ldots, x_n | \theta) \times e^{nc(\theta) + \theta T}$; where, $T = \sum_{i=1}^{n} t(x_i)$.

$$\frac{\partial \log L}{\partial\theta} = 0 \Rightarrow nc'(\theta) + T = 0$$

$$\Rightarrow -c'(\theta) = T/n$$

$$\Rightarrow \hat{\mu}_{MLE} = T/n.$$

Conjugate Prior:

$\pi(\theta) \propto \exp[mc(\theta) + \theta s]$; $\theta \in (a,b)$. We assume $\pi(a) = \pi(b) = 0$.

Posterior distribution:-

$\pi(\theta|\underset{\sim}{x}) \propto e^{(m+n)c(\theta) + \theta(s+T)}$.

[Another way of looking at conjugate prior is as follows:
Consider a prior $\pi(\theta) = 1$. Take m to be +ve integer and think of a hypothetical sample of size m with hypothetical data $x_1', \ldots, x_m'$ s.t. $\sum_{i=1}^{m} t(x_i') = s$ then

$$\pi(\theta | x_1', \ldots, x_m') \propto f(x_1', \ldots, x_m' | \theta)\pi(\theta)$$
$$= \exp\left\{\theta \cdot \sum_{i=1}^{m} t(x_i') + mc(\theta)\right\}$$
$$= e^{\theta s + mc(\theta)}.]$$

Prior mean:

$$E^{\pi}(\mu) = E^{\pi}\left(-c'(\theta)\right)$$
$$= -k\int_a^b c'(\theta)\, e^{mc(\theta) + \theta s}\, d\theta$$
$$= -\frac{k}{m}\int_a^b \frac{d}{d\theta}\left(e^{mc(\theta)}\right).e^{\theta s}\, d\theta$$
$$= -\frac{k}{m}\left[\left\{e^{\theta s} e^{mc(\theta)}\right\}_a^b - \int_a^b s e^{\theta s} e^{mc(\theta)}\, d\theta\right]$$
$$= -\frac{k}{m}\left[\left\{\pi(b) - \pi(a)\right\} - s\int_a^b \pi(\theta)\, d\theta\right]$$
$$= -\frac{1}{m}\left[\left\{\pi(b) - \pi(a)\right\} - s\right]$$
$$= \frac{s}{m}.$$

Similar calculations show that posterior mean of μ is

$$\frac{s+T}{m+n} = \underbrace{\frac{m}{m+n}\left(\frac{s}{m}\right)}_{\text{prior guess}} + \underbrace{\frac{n}{m+n}\left(\frac{T}{n}\right)}_{\text{classical estimate}}$$

So, the posterior mean is a weighted average of the prior guess $\frac{s}{m}$ and MLE $\frac{T}{n}$ and the weights are proportional to the precision parameter m and sample size n.

For multiparameter exponential families given by

$$p(\underset{\sim}{x}|\underset{\sim}{\theta}) = \exp\left\{c(\underset{\sim}{\theta}) + \sum_{i=1}^{d}\theta_i t_i(\underset{\sim}{x}) + h(\underset{\sim}{x})\right\}$$

The conjugate prior now takes the form

$$p(\underset{\sim}{\theta}) = \exp\left\{m A(\underset{\sim}{\theta}) + \sum_{i=1}^{d}\theta_i s_i\right\}$$

(4). $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Gamma (γ, θ) ; γ is known (>0).

Likelihood function:

$$f(\underset{\sim}{x}|\theta) = \prod_{i=1}^{n}\frac{\theta^\gamma}{\Gamma\gamma}x^{\gamma-1}e^{-\theta x_i} = \frac{\theta^{n\gamma}}{(\Gamma\gamma)^n}\left(\prod_{i=1}^{n}x_i\right)^{\gamma-1}e^{-\theta\sum_{i=1}^{n}x_i}$$

conjugate prior: $\pi(\theta) = \frac{\beta^\alpha}{\Gamma\alpha}\theta^{\alpha-1}e^{-\beta\theta}$; $\theta>0$

Posterior density: $\pi(\theta|\underset{\sim}{x}) \propto \theta^{\alpha+n\gamma-1}e^{-(\beta+\Sigma x_i)\theta}$,

a Gamma $\left(\alpha+n\gamma, \beta+\sum_{i=1}^{n}x_i\right)$ density.

Posterior mean and variance:

$$E(\theta|\underset{\sim}{x}) = \frac{\alpha+n\gamma}{\beta+\Sigma x_i}, \quad \text{Var}(\theta|\underset{\sim}{x}) = \frac{\alpha+n\gamma}{(\beta+\Sigma x_i)^2}.$$

(5) $f(x|\theta)$ is $N(\theta,1)$ density.

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\theta)^2\right)$$

$$= \exp\left(\theta x - \frac{1}{2}\theta^2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$$

So, $c(\theta) = -\frac{1}{2}$

$t(x) = x$

$T = \sum_{i=1}^{n}t(x_i) = n\bar{x}$

$\mu = E_\theta(X) = \theta = -c'(\theta)$.

$\pi(\theta) \propto \exp\left[-\frac{m}{2}\theta^2 + \theta s\right]$

$\propto \exp\left[-\frac{m}{2}\left(\theta - \frac{s}{m}\right)^2\right]$ which is $N\left(\frac{s}{m}, \frac{1}{m}\right)$.

Prior mean: $\frac{s}{m}$.

Posterior mean: $\frac{s+n\bar{x}}{m+n}$.

Predictive distribution:- Distribution of a future observation.

Let $X_1, \ldots, X_n$ be i.i.d. $f(x|\theta)$ and $\theta \sim \pi(\theta)$.

Prior predictive distribution:- Prior predictive distribution of a future observation X is given by density $m(x) = \int f(x|\theta)\pi(\theta)\,d\theta$.

Intuitively θ is known, X will follow $f(x|\theta)$, (H) we don't know θ but $\theta \sim \pi(\theta)$. Therefore, predictive distn is obtained as an average $f(x|\theta)$ with respect to $\pi(\theta)$.

Similarly, we can define prior predictive distn of first n observation as

$$m(x_1, x_2, \ldots, x_n) = \int_{(H)} f(x_1, \ldots, x_n|\theta)\,\pi(\theta)\,d\theta.$$

Posterior Predictive distribution:- Distribution of X_{n+1} given $X_1, \ldots, X_n$ is

$$f(x_{n+1}|x_1, \ldots, x_n) = \int_{(H)} f(x_{n+1}|\theta)\,\pi(\theta|x_1, \ldots, x_n)\,d\theta.$$

where $\pi(\theta|x_1, \ldots, x_n)$ is density of the posterior distn of θ given $x_1, \ldots, x_n$. Then $f(x_{n+1}|x_1 \ldots x_n)$ is called posterior predictive distn of X_{n+1} given $X_1, \ldots, X_n = (x_1, \ldots, x_n)$.

Example:- 1. $X_1, \ldots, X_n \sim \text{Ber}(\theta)$

$$f(x|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Consider $\pi(\theta) = 1$, then $\pi(\theta|\underset{\sim}{x}) \propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

$$f(x_{n+1}|x_1, \ldots, x_n) = \int_0^1 \theta^{x_{n+1}}(1-\theta)^{1-x_{n+1}}\,\pi(\theta|\underset{\sim}{x})\,d\theta$$

$$= \frac{\int_0^1 \theta^{\sum_1^{n+1} x_i}(1-\theta)^{n+1-\sum_1^{n+1} x_i}\,d\theta}{\text{Beta}\left(\sum_1^n x_i + 1,\ n - \sum_1^n x_i + 1\right)}$$

$$P(X_{n+1}=1 \mid X_1=1, \ldots, X_n=1) = \frac{\text{Beta}\left(\sum_1^{n+1} x_i + 1,\ n+2-\sum_1^{n+1} x_i\right)}{\text{Beta}\left(\sum_1^n x_i + 1,\ n - \sum_1^n x_i + 1\right)} = \frac{\sum_1^n x_i + 1}{n+2}$$

$$= \frac{n+1}{n+2}.$$

Example 2: (Normal Means)

$x_1, x_2, \ldots, x_n$ be iid $N(\theta, \sigma^2)$, σ^2 known.

$\theta \sim \pi(\cdot) = N(\mu, \tau^2)$, μ, τ^2 known.

$\pi(\theta | \underset{\sim}{x})$ = Normal with mean $= \dfrac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$ and

$$\text{variance} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

$$f(x_{n+1} | x_1, \ldots, x_n) = \int N(x_{n+1} | \theta, \sigma^2) \pi(\theta | \underset{\sim}{x}) \, d\theta$$

which is a normal distn. with

$$\text{mean} = E\left[X_{n+1} \middle| \underset{\sim}{x}\right] = E\left[E\left(X_{n+1} \middle| \underset{\sim}{x}, \theta\right)\right]$$

$$= E\left[\theta \middle| \underset{\sim}{x}\right]$$

$$\text{Var}\left(X_{n+1} \middle| \underset{\sim}{x}\right) = \text{Var}\left[E\left\{X_{n+1} \middle| \theta, \underset{\sim}{x}\right\}\right] + E\left[\text{Var}\left\{X_{n+1} \middle| \theta, \underset{\sim}{x}\right\}\right]$$

$$= \text{Var}\left(\theta \middle| \underset{\sim}{x}\right) + \sigma^2$$

$$= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \sigma^2.$$

Exercise:- For the above example, show that the prior predictive distn. (joint distn) of $X_1, X_2, \ldots, X_n$ is normal with $E(X_i) = \mu$ and $\text{Var}(X_i) = \tau^2 + \sigma^2 \; \forall \, i$ and $\text{Cov}(X_i, X_j) = \tau^2 \; \forall \; i \neq j$.

Prediction of a future observation by a single number $t(x_1, x_2, \ldots, x_n)$ based on $x_1, x_2, \ldots, x_n$ with squared error loss amounts to considering prediction loss $E\left[(X_{n+1} - t)^2 \middle| x_1, \ldots, x_n\right]$ which is minimum at $t = E(X_{n+1} | x_1, \ldots, x_n)$.

Multiparameter models :-

Ex.1. (Normal)

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$; both μ, σ^2 unknown.

A noninformative prior for μ and σ^2 is

$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$ (conventional improper prior uniform on $(\mu, \log \sigma^2)$)

Joint posterior density

$$\pi(\mu, \sigma^2 | \underset{\sim}{x}) \propto (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\{(n-1)s^2 + n(\mu-\bar{x})^2\}\right],$$

where $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$.

Conditional posterior density of μ given $(\sigma^2, \underset{\sim}{x})$ is given by

$$\pi(\mu | \sigma^2, \underset{\sim}{x}) \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right).$$

Marginal posterior distn

$$\pi(\sigma^2 | \underset{\sim}{x}) \propto \int (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\{(n-1)s^2 + n(\mu-\bar{x})^2\}\right] d\mu$$

$$\propto (\sigma^2)^{-\frac{(n+1)}{2}} \cdot \exp\left[-\frac{1}{2\sigma^2}(n-1)s^2\right]$$

This is a scale inverse $= \chi^2(n-1, s^2)$ density.

[Putting : $t = \frac{(n-1)s^2}{\sigma^2}$, $\sigma^2 = \frac{(n-1)s^2}{t}$, $\frac{d\sigma^2}{dt} = -\frac{(n-1)s^4}{t^2}$]

$$\pi(t | \underset{\sim}{x}) \propto \left[\frac{(n-1)s^2}{t}\right]^{-\left(\frac{n+1}{2}\right)} \exp\left[-t/2\right] \frac{(n-1)s^2}{t^2}$$

$$\propto t^{\frac{n-1}{2}-1} \exp\left[-t/2\right]$$

Conditional distn of $\frac{(n-1)s^2}{\sigma^2}$ given $\underset{\sim}{x}$ is χ^2_{n-1}.

[Sampling theory result :-
Distn of $\frac{(n-1)s^2}{\sigma^2}$ (given μ, σ^2) is χ^2_{n-1}.]

Marginal Posterior distribution of μ:-

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

$$\propto \int_0^\infty (\sigma^2)^{-n/2-1} e^{-\frac{A}{2\sigma^2}} d\sigma^2$$

$$\left[A = (n-1)s^2 + n(\mu-\bar{x})^2 \right.$$
$$\left. z = \frac{A}{2\sigma^2}, \quad \sigma^2 = \frac{A}{2z} \right]$$

$$= \int_0^\infty \left(\frac{A}{2z}\right)^{-n/2-1} e^{-z} \frac{A}{2z^2} dz$$

$$\propto A^{-n/2} \int_0^\infty z^{n/2-1} e^{-z} dz$$

$$\propto \left[1 + \frac{n(\mu-\bar{x})^2}{(n-1)s^2}\right]^{-n/2}$$

This is the t-distribution $t_{n-1}\left(\bar{x}, \frac{s^2}{n}\right)$.

[$\theta \sim t_\gamma(\mu, \sigma^2)$ mean $p(\theta)$ = constant]

Standard $t \sim t_\gamma(0,1)$ denoted by t_γ.

Note that $\frac{\mu-\bar{x}}{s/\sqrt{n}}$ (given $\underset{\sim}{x}$) $\sim t_{n-1}$.

[Sampling theory result: $\frac{\mu-\bar{x}}{s/\sqrt{n}}$ (given μ, σ^2) $\sim t_{n-1}$]

Multiparameter Models :-

Example (Normal Data with a conjugate prior distribution) :-

$x_1, x_2, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$; μ, σ^2 both unknown.

Prior: $\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\,\pi(\sigma^2)$

$\mu|\sigma^2 \sim \pi(\mu|\sigma^2) = N\left(\mu_0, \frac{\sigma^2}{k_0}\right)$

$\sigma^2 \sim \pi(\sigma^2) = \text{Inverse-}\chi^2(\gamma_0, \sigma_0^2)$ $\left[\text{d.f.} = \gamma_0, \text{scale} = \sigma_0^2\right]$

Then $\pi(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{1/2}} \exp\left[-\frac{k_0}{2\sigma^2}(\mu-\mu_0)^2\right]$

$\times (\sigma^2)^{-\left(\frac{\gamma_0}{2}+1\right)} e^{-\gamma_0\sigma_0^2/2\sigma^2}$

We call this the $N\text{-Inv-}\chi^2\left(\mu_0, \frac{\sigma^2}{k_0}; \gamma_0, \sigma_0^2\right)$ density.

Joint Posterior Density :-

$$\pi(\mu, \sigma^2|\underset{\sim}{x}) \propto (\sigma^2)^{-\frac{(\gamma_0+3)}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{\gamma_0\sigma_0^2 + k_0(\mu-\mu_0)^2\right\}\right]$$

$$\times \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\mu-\bar{x})^2\right\}\right]$$

$$= N\text{-Inv-}\chi^2\left(\mu_n + \frac{\sigma^2}{k_n}; \gamma_n, \sigma_n^2\right)$$

where, $\mu_n = \frac{k_0}{k_0+n}\mu_0 + \frac{n\bar{x}}{k_0+n}$; $k_n = k_0 + n$; $\gamma_n = \gamma_0 + n$.

$\gamma_n\sigma_n^2 = \gamma_0\sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_0+n}(\bar{x}-\mu_0)^2$ (check)

Conditional Posterior Density :-

$\pi(\mu|\sigma^2, \underset{\sim}{x})$: $\mu|\sigma^2, \underset{\sim}{x} \sim N\left(\mu_n, \frac{\sigma^2}{k_n}\right)$

Marginal Posterior Densities :- $\pi(\sigma^2|\underset{\sim}{x})$ and $\pi(\mu|\underset{\sim}{x})$:

$\pi(\sigma^2|\underset{\sim}{x})$ is an $\text{Inv-}\chi^2(\gamma_0, \sigma_n^2)$ density.

$$\pi(\mu|\underset{\sim}{x}) \propto \left[1 + \frac{k_n(\mu-\mu_0)}{\gamma_n\sigma_n^2}\right]^{-(\gamma_n+1)/2}$$ (check)

is a t-distribution with location $= \mu_n$, scale $= \frac{\sigma_n}{\sqrt{k_n}}$ and degree of freedom $= \gamma_n$.

The Multivariate Normal Model:-

$x_1, x_2, \ldots, x_n$ iid $N_p(\mu, \Sigma)$; Σ is p.d.

$$f(x_1, x_2, \ldots, x_n | \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]$$

Consider Σ is known

Conjugate prior distn. for μ is : $\mu \sim N(\mu_0, \Lambda_0)$; μ_0, Λ_0 known.

Posterior density of μ is :

$$\pi(\mu | \underset{\sim}{x}, \Sigma) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu) - \frac{1}{2}(\mu-\mu_0)'\Lambda_0^{-1}(\mu-\mu_0)\right]$$

Now,

$$\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu) + (\mu-\mu_0)'\Lambda^{-1}(\mu-\mu_0)$$

$$= \sum_{i=1}^{n}(x_i-\bar{x})'\Sigma^{-1}(x_i-\bar{x}) + n(\mu-\bar{x})'\Sigma^{-1}(\mu-\bar{x}) + (\mu-\mu_0)'\Lambda^{-1}(\mu-\mu_0)$$

$$= \mu'\left(n\Sigma^{-1}+\Lambda_0^{-1}\right)\mu - 2\mu'\left(n\Sigma^{-1}\bar{x}+\Lambda_0^{-1}\mu_0\right) + \text{constant (free of } \mu)$$

$$= (\mu-\mu_n)'\Lambda_n^{-1}(\mu-\mu_n)$$

where, $\mu_n = \left(\Lambda_0^{-1}+n\Sigma^{-1}\right)^{-1}\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}\right)$

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}.$$

Then $\pi(\mu | \underset{\sim}{x}, \Sigma) \propto \exp\left[-\frac{1}{2}(\mu-\mu_n)'\Lambda_n^{-1}(\mu-\mu_n)\right]$

$= N(\mu_n, \Lambda_n)$ density.

The parameter mean μ_n is a weighted average of $\bar{x}$ and the prior mean μ_0 with weights given by data ($\bar{x}$) and prior precision matrices.

Consider now the case with unknown Σ

Conjugate prior distn. for (μ, Σ) is normal-inserve Wishart $\left(\mu_0, \frac{\Sigma_0}{k_0}; \gamma_0, \Lambda_0\right)$.

So, $\mu|\Sigma \sim N\left(\mu_0, \frac{\Sigma}{k_0}\right)$, $\Sigma \sim$ Inverse-Wishart$_{\gamma_0}\left(\Lambda_0^{-1}\right)$.

[If $W^{-1} \sim$ Wishart$_\gamma(S)$ (γ = d.f., S = $p \times p$ scale matrix)
Then $W \sim$ Inserve-Wishart$_\gamma(S^{-1})$.

$$f(W) = \left(2^{\gamma p/2} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(\frac{\gamma+1-i}{2}\right)\right)^{-1} \times |S|^{\gamma/2} |W|^{-(\gamma+p+1)/2} \times \exp\left[-\frac{1}{2} \text{tr}\left(SW^{-1}\right)\right]; \; W \text{ is p.d.}$$

The joint prior density is:

$$\pi(\mu, \Sigma) \propto |\Sigma|^{-\left(\frac{\gamma_0+p}{2}+1\right)} \exp\left[-\frac{1}{2}\text{tr}\left(\Lambda_0\Sigma^{-1}\right) - \frac{k_0}{2}(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0)\right]$$

The posterior density is $\pi(\mu, \Sigma|\underset{\sim}{X})$ = a Normal-Inverse Wishart density with parameters

$$\mu_n = \frac{k_0}{k_0+n}\mu_0 + \frac{n}{k_0+n}\bar{X},$$

$$k_n = k_0 + n, \quad \gamma_n = \gamma_0 + n,$$

$$\Lambda_n = \Lambda_0 + S + \frac{k_0 n}{k_0+n}(\bar{X}-\mu_0)(\bar{X}-\mu_0)'$$

where, S is the correlated sum of saware matrix, i.e.,

$$S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'.$$

The marginal posterior distn. of μ is a multivariate t-distn.

Regression Model:- $y_i, x_{i1}, \ldots, x_{ip}$; $i = 1, 2, \ldots, n$.

$\underset{\sim}{y} = (y_1, \ldots, y_n)'$, $X = ((x_{ij}))_{n \times p}$

$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$; where, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

$E(y_i | \beta, X) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

$y | \beta, \sigma^2 \sim f(y | \beta, \sigma^2) = N_n(X\beta, \sigma^2 I)$ [Taking $y = \underset{\sim}{y}$]

Consider the standard non-informative prior $\pi(\beta, \sigma^2)$; where,

$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$; $f(y | \beta, \sigma^2)$: likelihood.

$$\pi(\beta, \sigma^2 | y) \propto f(y | \beta, \sigma^2)\pi(\beta, \sigma^2)$$

$$= \frac{1}{(\sigma^2)^{n/2+1}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right]$$

$$= (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\left\{(\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) + \underbrace{(y - X\hat{\beta})'(y - X\hat{\beta})}_{(n-p)s^2, \text{ say}}\right\}\right]$$

where, $\hat{\beta} = (X'X)^{-1}X'y$.

$$\pi(\beta | \sigma^2, y) \propto f(y | \beta, \sigma^2)\pi(\beta)$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'V^{-1}(\beta - \hat{\beta})\right] \text{ ; where } V = (X'X)^{-1}$$

which is $N_p(\hat{\beta}, \sigma^2 V)$.

$$\pi(\sigma^2 | y) = \int \pi(\beta, \sigma^2 | y)\, d\beta$$

$$= \text{constant} \times (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}(y - X\hat{\beta})'(y - X\hat{\beta})\right]$$

$$\times \left|(X'X)^{-1}\sigma^2\right|^{1/2}$$

$$\propto (\sigma^2)^{-\frac{n-p}{2}-1} \exp\left[-\frac{1}{2\sigma^2}(n-p)s^2\right]$$

which is a scaled inverse $\chi^2(n-p, s^2)$ distribution;

where, $s^2 = \frac{1}{n-p}(\underset{\sim}{y} - X\hat{\beta})'(\underset{\sim}{y} - X\hat{\beta})$.

Posterior density of γ is given by:

$$\pi(\beta|y) = \int \pi(\beta, \sigma^2|y)\, d\sigma^2$$

$$= \text{constant} \times \int (\sigma^2)^{-n/2-1} \exp\left(-\frac{A}{2\sigma^2}\right) d\sigma^2$$

where, $A = (n-p)s^2 + (\beta-\hat{\beta})(X'X)(\beta-\hat{\beta})$

$$\propto A^{-n/2} \quad [\text{As obtained earlier in } N(\mu, \sigma^2) \text{ example}]$$

$$\propto \left[1 + \frac{(\beta-\hat{\beta})X'X(\beta-\hat{\beta})}{(n-p)s^2}\right]^{-n/2}$$

which is a multivariate t-distn. with $(n-p)$ d.f.

Location vector : $\hat{\beta}$, scale matrix : $s^2(X'X)^{-1}$.

The Multinomial Model:- This model is used to describe data for which each observation is one of k possible outcomes.

Sample space $= \{1, 2, \ldots, k\}$, say.

$x_1, x_2, \ldots, x_n$ iid with $P(x_i = j) = p_j$; $j = 1, 2, \ldots, k$.

$p_i \geq 0 \ \forall j$ and $\sum_{j=1}^{k} p_j = 1$.

Likelihood function: $f(x_1, \ldots, x_n|p) = p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$; n_j: # of x_i's equal to j.

Conjugate prior:- The conjugate prior distn. is Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$; [Multivariate generalisation of Beta]

$\alpha_j > 0 \ \forall j$.

$$\pi(p_1, \ldots, p_k | \alpha_1, \ldots, \alpha_k)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \ldots p_k^{\alpha_k-1} \quad ; \ p_j \geq 0 \ \forall j \text{ and } \sum_{j=1}^{k} p_j = 1.$$

Posterior distn. of p is Dirichlet$(\alpha_1+n_1, \ldots, \alpha_k+n_k)$.

Note that $\alpha_j = 1 \ \forall j$ corresponds to a uniform distribution.

Bayesian Nonparametrics:- P is a random probability distn.
We need a prior prob. distn of the "random" element P.
Challenges of Bayesian nonparametrics:-
Construction of prior
Posterior computation

Example (Bernoulli revisited): $\mathcal{X} = \{0, 1\}$
$X_1, X_2, \ldots, X_n$ iid $\text{Bin}(1, p)$.
$p^* = (1-p, p)$ where $p \sim \text{Beta}(\alpha_0, \alpha_1)$

$$E(p^*) = \left(E(1-p), E(p)\right) = \left(\frac{\alpha_1}{\alpha_0+\alpha_1}, \frac{\alpha_0}{\alpha_0+\alpha_1}\right)$$

$\alpha(\cdot)$ = measure of $\mathcal{X}$; i.e., $\bar{\alpha} = \dfrac{\alpha(\cdot)}{\alpha(\mathcal{X})}$

So, the prob. distn $\bar{\alpha}$ is the prior guess for random prob. distn P.

Definition:- A random measure P on $(\mathbb{R}, \mathcal{B})$ is said to possess a Dirichlet process distn $DP(\alpha)$ with base measure α, if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathcal{X}$ $(= \mathbb{R})$, $\left(P(A_1), \ldots, P(A_k)\right) \sim \text{Dirichlet}\left(\alpha(A_1), \ldots, \alpha(A_k)\right)$.

Proposition: If $P \sim DP(\alpha)$, then for any measurable set A and B

$$E(P(A)) = \bar{\alpha}(A)$$

$$V(P(A)) = \frac{\bar{\alpha}(A)\,\bar{\alpha}(A^c)}{1+\alpha(\mathcal{X})}$$

$$\text{Cov}(P(A), P(B)) = \frac{\bar{\alpha}(A \cap B) - \bar{\alpha}(A)\,\bar{\alpha}(B)}{1+\alpha(\mathcal{X})}$$

Theorem:- The posterior distn given a random sample from P where $P \sim DP(\alpha)$ is $DP\left(\alpha + \sum_{i=1}^{n} \delta_{x_i}\right)$; where δ_c is a measure degenerate at c.

Note: In real life problems, dirichlet mixture distn is highly useful to model the data.

Consistency of Posterior Distribution:-

$\underset{\sim}{X}_n = (X_1, \ldots, X_n) \sim f(\underset{\sim}{x}_n | \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$.

$\pi(\theta)$ is prior density.

$\pi(\theta | \underset{\sim}{X}_n)$ is posterior density.

$\pi(\cdot | \underset{\sim}{X}_n)$ is the corresponding posterior distn.

- **Definition:-** The sequence of the posterior distn $\pi(\cdot | \underset{\sim}{X}_n)$ is said to be consistent at some $\theta_0 \in \Theta$ if for every neighbourhood U of θ_0, $\pi(U | \underset{\sim}{X}_n) \to 1$ as $n \to \infty$ with prob. 1 w.r.t. the distn under θ_0.

- For a real parameter θ, consistency at θ_0 can be proved by showing $E(\theta | \underset{\sim}{X}_n) \to \theta_0$ and $\text{Var}(\theta | \underset{\sim}{X}_n) \to 0$ w.p.1.

- **Example:** $X_1, \ldots, X_n$ be iid Bernoulli(θ); $\theta \sim \text{Beta}(\alpha, \beta)$ (prior)

Posterior distn. is $\text{Beta}\left(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta\right)$ with

$$E(\theta | \underset{\sim}{X}_n) = \frac{\sum_{i=1}^n X_i + \alpha}{n + \alpha + \beta} \quad \text{and} \quad \text{Var}(\theta | \underset{\sim}{X}_n) = \frac{\left(\sum_{i=1}^n X_i + \alpha\right)\left(n - \sum_{i=1}^n X_i + \beta\right)}{(\alpha + \beta + n)^2 (\alpha + \beta + n + 1)}$$

As, $\frac{1}{n} \sum_{i=1}^n X_i \to \theta_0$ w.p.1 under θ_0 (SLLN)

It follows that $E(\theta | \underset{\sim}{X}_n) \to \theta_0$ and $\text{Var}(\theta | \underset{\sim}{X}_n) \to 0$ w.p.1 under θ_0

This implies that the posterior is consistent.

An important result related to consistency is the robustness of posterior inference w.r.t. choice of prior.

Let $X_1, \ldots, X_n$ be iid and π_1 and π_2 are two prior densities which are positive and continuous at θ_0, an interior point of Θ such that the corresponding posterior distn $\pi_1(\cdot | \underset{\sim}{X}_n)$ and $\pi_2(\cdot | \underset{\sim}{X}_n)$ are both consistent at θ_0,

$$\int_\Theta \left| \pi_1(\theta | \underset{\sim}{X}_n) - \pi_2(\theta | \underset{\sim}{X}_n) \right| d\theta \to 0$$

$$\Leftrightarrow \sup_A \left| \pi_1(A | \underset{\sim}{X}_n) - \pi_2(A | \underset{\sim}{X}_n) \right| \to 0$$

On choice of Priors :- Given data, a Bayesian needs the likelihood $p(x|\theta)$ and the prior $\pi(\theta)$. For many standard problems, the likelihood is known either from past experience or by conventions. To derive the Bayesian engine, one would need an appropriate prior.

Ideally one wants to choose a prior or a class of priors reflecting one's prior knowledge and belief about unknown parameters or about different hypotheses. This is a subjective choice. Choice of subjective priors, usually called elicitation of priors, is still rather difficult. We will discuss it in details later. But at this moment, let us see why it is different. Empirical studies have shown that experience and maturity helps the person in quantifying uncertainty about an event in the form of a probability. However assigning a fully specified prob. distn. to an unknown parameter is difficult even when the parameter has a physical meaning like length and, or, breadth of some article. In such cases, it may be realistic to expect elicitation of prior mean or variance to some other prior quantities but not a full specification of the distribution.

It is much more common to use what are called objective priors. When very little prior information is available, objective priors are also called non-informative prior. The term "non-informative prior" is no longer in favour among Bayesian because a complete lack of information is hard to define. There are standard algorithms to the construction of priors with zero information. Objective priors are typically improper but have proper posterior distribution. They are suitable for estimation problems and also for testing problems where both null and alternative hypothesis have the same distributions. The objective prior need to be suitably modified for sharp null hypothesis.

Different methods for construction of Objective Priors:-

We may do one of the following things to construct objective prior under general regularity condition:

1. Define a uniform distn. that takes into account the geometry of parameter space.
2. Minimize a suitable measure of information in the prior.
3. Choose a prior with same form of frequentist ideas because a prior with little information should lead to inference that is similar to frequentist inference.

- To fully define these methods, we have to specify the geometry in (1), the measure of information (2) and the frequentist ideas that are to be defined in (3). It is striking fact that (1) and (2) both lead to Jeffrey's prior, namely, $\pi(\theta) = \sqrt{|I(\theta)|}$, where, $I(\theta)$ is the Fisher Information Matrix. In the one dimensional case (3) leads to Jeffrey's prior.

- It may be noted that some statistical models posseses additional structure. Some are exponential families, some are location-scale families or more generally families invariant under a group of transformation.

- For each of these special classes, there is a different choice of objective priors. The objective priors for the exponential families came from the class of conjugate priors. In the class of location-scale families with scale parameter σ; the common objective prior is so called right invariant Haar measure, $\pi_1(\mu, \sigma) = 1/\sigma$ and the Jeffrey's prior is the left invariant Haar measure $\pi_2(\mu, \sigma) = \frac{1}{\sigma^2}$. There are strong reasons for preferring π_1 to π_2.

The Uniform Prior and its criticism:-

The first objective prior ever to be used is the uniform distribution over a compact interval. A common argument based on "ignorance" seems to have been that if we know nothing about θ why should I attach more density to one point than another. The principle of ignorance has been criticized by many. Essentially, the criticism is based on an invariance argument.

Let $\eta = \psi(\theta)$ be a one-to-one function of θ. If we know nothing about θ, then we know nothing about η also. So the principle of ignorance applied to η will imply our prior η is uniform; just as it had led to a uniform prior for θ. But this leads to contradiction.

To see this suppose ψ is differentiable and $p(\eta) = c$ on ψ (H) then the prior $p^*(\theta)$ for θ is $p^*(\theta) = p(\eta)|\psi'(\theta)|$

i.e., $p^*(\theta) = c\,|\psi'(\theta)|$ which is not a constant in general.

This argument also leads to an invariance principle. Suppose we have an algorithm that produces non-informative priors for both θ and η. Then these priors $p^*(\theta)$ and $p(\eta)$ should be connected by the equation;

$$p^*(\theta) = p(\eta)\,|\psi'(\theta)|$$

i.e., a non-informative prior should be invariant under one-to-one differentiable transformations.

Some arguments in favour of Jeffrey's prior as a non-informative Prior:-

Ghoshal, Ghosh, Ramamurthy (1997) constructed a uniform distn. taking into account the topology of the parameter space and showed that the uniform distn. is the Jeffrey's prior which automatically satisfy the invariance requirement.

In their construction one takes a compact subset of the parameter space and approximates this by a finite set of points in the Hellinger metric

$$d(P_\theta, P_{\theta'}) = \left[\int \left(\sqrt{p_\theta} - \sqrt{p_{\theta'}} \right)^2 dx \right]^{1/2} ;$$

where, p_θ and $p_{\theta'}$ are the densities of P_θ and $P_{\theta'}$. One thus puts a discrete uniform distn. on the approximating finite set of points and let the degree of approximation tend to zero. then the corresponding discrete uniform converge to Jeffrey's prior.

K is a compact metric space with a metric ρ. $S \subset K$ is called ϵ-dispersed if $\rho(x,y) \geq \epsilon \quad \forall \quad x, y \in S, x \neq y$. A maximal ϵ-dispersed set will be called an ϵ-net. An ϵ-net with maximum possible cardinality is said to be an ϵ Lattice is called the packing member (or ϵ-capacity) of K and is denoted by $D(\epsilon, K)$.

For $X \subset K$ define $P_\epsilon(X) = \dfrac{D(\epsilon, X)}{D(\epsilon, K)}$.

In the paper, in place of ρ one takes the hellinger metric then taking a compact subset K of Θ one has for $\Omega \subset K$

$$\lim_{\epsilon \to 0} \frac{D(\epsilon, \Omega)}{D(\epsilon, K)} = \frac{\int_\Omega \sqrt{\det I(\theta)}\, d\theta}{\int_K \sqrt{\det I(\theta)}\, d\theta} .$$

Jeffrey's Prior as a Probability matching prior :-

One should expect an objective prior with low information to provide inference similar to that based on uniform prior for θ in $N(\theta, 1)$, where the Bayesian and frequentist's answers were exactly identical. It may be recalled that in this problem, the posterior distn. of $\theta - \bar{X}$ given $\underset{\sim}{X}$ is identical with the frequentist distn. of $\theta - \bar{X}$ given θ.

In general case, we will not get exactly the same distn. but only up to $O(1/n)$. We give a precise definition of probability matching prior for a single parameter below.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $p(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}$. Assume regularity condition of the parameter with normal $\left(\hat{\theta}, (nI(\hat{\theta}))^{-1}\right)$ as the leading term. For $0 < \alpha < 1$, choose $\theta_\alpha(X)$ depending on the prior $p(\theta)$ s.t.

$$P\left(\theta \le \theta_\alpha(X) \mid X\right) = 1 - \alpha + O_p(1/n).$$

We say $p(\theta)$ is probability matching (to the 1st order) if

$$P_\theta\left(\theta \le \theta_\alpha(X)\right) = 1 - \alpha + O(1/n)$$

Uniformly on compact set of θ using normally approximation,

$$P\left(\theta < \hat{\theta} + \frac{z_\alpha}{\sqrt{n}} \mid X\right) = P_\theta\left(\theta < \hat{\theta} + \frac{z_\alpha}{\sqrt{n}}\right) + O_p(1/n).$$

under θ (uniformly on compact subsets of θ).

Probability matching priors :- Ref: Dutta & Mukherjee (2004).

In this approach, a common goal is to find matching prior for posterior quantities, we want

$$P\left(\theta < \theta^{(1-\alpha)}(\pi, X)\right) = 1 - \alpha + O\left(n^{-1/2}\right)$$

for each α; $0 < \alpha < 1$, where $\theta^{(1-\alpha)}(\pi, X)$ is the $(1-\alpha)$th posterior quantile of θ under prior $\pi(\,)$ given data X.

In order to do this first one finds an expression for the posterior density of $\sqrt{n}(\theta - \hat{\theta})$.

$$\pi(\theta|X) = \frac{\pi(\theta)\exp(nl(\theta))}{\int \pi(\theta)\exp(nl(\theta))\,d\theta}$$

$$= \frac{\pi(\theta)\exp\left(n\left(l(\theta) - l(\hat{\theta})\right)\right)}{\int \pi(\theta)\exp\left(n\left(l(\theta) - l(\hat{\theta})\right)\right)d\theta}$$

Here $l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(f(x_i, \theta))$

Let $h = \sqrt{n}(\theta - \hat{\theta})$

$$\pi^*(h|X) = \frac{b(h,X)}{\int b(h,X)\,dh} \quad ; \text{where,}$$

$$b(h,X) = \pi\left(\hat{\theta} + \frac{h}{\sqrt{n}}\right)\exp\left\{n\left(l\left(\hat{\theta} + \frac{h}{\sqrt{n}}\right) - l(\hat{\theta})\right)\right\}$$

Then do Taylor expansion about $\hat{\theta}$ in both these terms:

This will finally give

$$\pi^*(h|X) \ldots\ldots \phi_h(0, c^{-1})\left[1 + \frac{1}{\sqrt{n}}g_1(h, \underset{\sim}{X}) + \frac{1}{n}g_2(h, \underset{\sim}{X})\right] + O(1/n);$$

where, $\phi_h(0, c^{-1})$ is the density of a normal distn. with mean zero and variance as the inverse of per observation observed Fisher Information matrix (c), evaluated at h.

Then in the next step one finds an expression for $\theta^{(1-\alpha)}(\pi, X)$, the $(1-\alpha)$th posterior quantile of θ under prior $\pi(\cdot)$. This can be obtained by using the expansion of the posterior already estimated. In general, it will be of the form:

$$\theta^{(1-\alpha)}(\pi, X) = \hat{\theta} + \frac{1}{\sqrt{n}}c(\underset{\sim}{X}) + \left[z_\alpha + \frac{1}{\sqrt{n}}(\;) + \frac{1}{n}(\;)\right]$$

The last and very crucial step is to calculate an expression for the frequentist coverage probability $P_\theta\left(\theta \le \theta_1^{(1-\alpha)}(\pi, X)\right)$

$$P\left(\theta \le \theta_1^{(1-\alpha)}(\pi, X) \,|\, X\right) = 1 - \alpha + O(1/n)$$

Using the shrinkage arguement directed by J.K. Ghosh one can find an expression for this quantity. For details, see Dutta and Mukherjee (2004). Using this technique, one ultimately gets —

$$P_\theta\left(\theta \le \theta^{(1-\alpha)}(\pi, X)\right) = (1-\alpha) + \frac{1}{\sqrt{n}} \cdot \frac{\phi(z_\alpha)}{\pi(\theta)} A_1(\pi, \theta) + \frac{1}{n} \cdot \frac{z_\alpha \phi(z_\alpha)}{\pi(\theta)} A_2(\pi, \theta) + O(1/n),$$

for specific function A_1 and A_2.

In this case, $A_1(\pi, \theta) = \frac{d}{d\theta}\left\{\frac{\pi(\theta)}{I^{1/2}}\right\}$

So, for a prior to be first order probability matching, one needs to have $A_1(\pi, \theta) = 0$; i.e., $\frac{d}{d\theta}\left\{\frac{\pi(\theta)}{I^{1/2}}\right\} = 0$.

The solution of this equation comes out as

$\pi(\theta) \propto \sqrt{I(\theta)}$, the well known Jeffry's prior.

Jeffrey's prior as a minimizer of information :-

The amount of information to be expected from an experiment about some quantities of interest naturally depends on the available prior knowledge ; the more prior information available, the less information can be learned from the data. An infinitely large experiment would eventually provide all missing information, thus it is possible to obtain a measure of the amount of missing information as a limiting form of a function of the prior distn. Using an idea of Lindley, Bernardo suggested that a Kullback - Leibdler divergence between prior and posterior namely

$$J(P(\theta), X) = E\left(\log \frac{P(\theta|X)}{P(\theta)}\right)$$ as the amount of information which may be expected to be provided by an experiment about the value of θ. It is easy to see that when the prior $P(\theta)$ is nearly degenerate at a point, so will be the posterior, then J will be near 0.

On the other hand, if $P(\theta)$ is rather diffuse $P(\theta|\underset{\sim}{x})$ will differ a lot from $P(\theta)$ at least for large or moderate n because $P(\theta|x)$ would be approximately normal with mean $\hat{\theta}$ and variance of the order $O(1/n)$. The substantial difference in prior and posterior will be reflected in a large value for J. To sum up: J is small when $P(\theta)$ is nearly degenerate and large when P is diffuse, i.e., J combines how diffuse is the prior. Therefore it makes sense to maximize J w.r.t. the prior $P(\theta)$.

Bernardo suggested one should not work with sample size n of the given data and maximize J for this n. For one thing, this would be technically forbidding in most cases and more importantly, the functional J is expected to be a nice function of the prior only assymptotically.

We will do the maximisation assymptotically. One may approximate J by $\hat{J}$ by using idea of posterior normality, i.e., by replacing $p(\theta|x)$ by $\hat{p}(\theta|x)$ in the definition of J, $\hat{p}(\theta|x)$ being the density of the approximately normal distn..

With this approximation, one gets

$$\hat{J} = \left\{ -\frac{d}{2}\log(2\pi) - \frac{d}{2} + \frac{d}{2}\log n \right\} + \int \log\left(\sqrt{I(\theta)}\cdot P_i(\theta)\right) d\theta - \int_{K_i} \log\left(P_i(\theta)\right) P_i(\theta)\, d\theta + {}^{K_i} O_p(1).$$

$(-J)$ is a measure of information in a prior.

To maximize J w.r.t. $p(\theta)$; we will try assymptotic maximization of J.

Towards that ; let us concentrate on a compact subset K_i of (H), you can show that

$$J(p(\theta), X) = \int_{(H)} \left\{ \int_X \left[\int_{(H')} \frac{\log p(\theta'|x)}{p(\theta')} p(\theta'|x)\, d\theta' \right] p(x|\theta)\, dx \right\} p(\theta)\, d\theta$$

We will first assume that

$$J(p(\theta), X) \approx \hat{J}(\hat{p}(\theta), X)$$

We will assume that J can be approximated well approximately by replacing $p(\theta'|x)$ by $\hat{p}(\theta'|x)$; where $\hat{p}(\theta'|x)$ is the density of a $N\left(\hat{\theta}, \frac{1}{nI(\theta)}\right)$ distn. evaluated at θ'.

Restricted to compact set K_i

$$\hat{J}(\hat{p}(\theta), X) = \int_{K_i} \left\{ \int_X \left\{ \int_{\mathbb{R}} \log \frac{\hat{p}_i(\theta'|x)}{p_i(\theta')} \hat{p}_i(\theta'|x)\, d\theta' \right\} p(x|\theta)\, dx \right\} p_i(\theta)\, d\theta .$$

$$\theta' \approx N\left(\hat{\theta}, \frac{1}{nI(\hat{\theta})}\right).$$

$$\hat{\phi}_i(\theta'|x) = \frac{\sqrt{nI(\hat{\theta})}}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} nI(\hat{\theta})(\theta' - \hat{\theta})^2\right)$$

$$\log \hat{p}_i(\theta'|x) = \tfrac{1}{2}\log n + \tfrac{1}{2}\log I(\hat{\theta}) - \tfrac{1}{2}\log 2\pi - \tfrac{1}{2} nI(\hat{\theta})(\theta' - \hat{\theta})^2$$

$\hat{J}(\hat{p}(\theta), X)$ reduces to

$$\simeq \int_{K_i} \left[\int_X -\log p_i(\hat{\theta}) p(\underset{\sim}{x}|\theta)\, dx \right] p_i(\theta)\, d\theta$$

$$\simeq -\int_{K_i} \log\left(p_i(\hat{\theta})\right) p_i(\theta)\, d\theta$$

Big integral $\simeq \left(-\tfrac{1}{2}\log 2\pi - 1/2 + 1/2 \log n\right) + \int_{K_i} \log\left(\sqrt{I(\theta)}\right) p_i(\theta)\, d\theta$

$$- \int_{K_i} \log\left(p_i(\theta)\right) p_i(\theta)\, d\theta$$

$$= c + \int_{K_i} \log\left(\frac{\sqrt{I(\theta)}}{p_i(\theta)}\right) p_i(\theta)\, d\theta$$

The prior maximizes this quantity if

$$p_i(\theta) = \begin{cases} c_i \sqrt{I(\theta)} & \text{if } \theta \in K_i \\ 0 & \text{ow} \end{cases}$$

where, c_i is a normalising constant s.t. $c_i \sqrt{I(\theta)}$ is a prob. density on K_i.

Bayesian asymptotics: (Posterior consistency and Posterior normality)

The Bernstein-Von Mises theorem about posterior on normality says the following:

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$; $\theta \in \Theta \subset \mathbb{R}$

where, Θ is an open subset of $\mathbb{R}$.

Let $\theta_0 \in \Theta$; where Θ is an open subset of $\mathbb{R}$.

Let $\hat{\theta}_n$ be a strongly consistent solution of the likelihood equation. Assume all the usual regularity condition needed for proving assymptotic normality of consistent roots of MLE.

(i) for any $\delta > 0$

$$P_{\theta_0}\left[\text{for some } \epsilon > 0, \sup_{|\theta-\theta_0|>\epsilon} \frac{1}{n}\left[l_n(\theta) - l_n(\theta_0)\right] < -\epsilon, \text{ for all sufficiently large } n\right] = 1.$$

(ii) The prior has a density $\pi(\theta)$ w.r.t. lesbeague measure, which is continuous and positive at θ_0.

Let $\pi_n^*(t|X_1, X_2, \ldots, X_n)$ be the posterior density of $t = \sqrt{n}(\theta - \hat{\theta}_n$

Then $\lim_{n\to\infty} \int \left| \pi_n(t|X_1, \ldots, X_n) - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt$

$= 0$ with probability 1, under P_{θ_0}.

Bernestien - Van - Mises Theorem :- A standard result in the asymptotic theory of MLE on its asymptotic normality. Its Bayesian parallel is a result often referred to as Bernestien - Van - Misen theorem – a result on the asymptotic normality of the posterior distn. We consider the case of a real parameter θ, $\theta \in \Theta \subset \mathbb{R}$, where Θ is open.

For simplicity, let us take $\Theta = \mathbb{R}$. But the result goes through exactly as stated below even though Θ is an open interval in $\mathbb{R}$.

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \in \mathbb{R}$.

P_θ has the density $f(x|\theta)$. We make the following regularity assumptions. Fix $\theta_0 \in \Theta$ (regarded as the true value of the parameter)

Notation : $l(\theta, x) = \log f(x|\theta)$

$$L_n(\theta) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} l(\theta, x_i).$$

The assumptions are as follows :

(A1) The set $\{x : f(x|\theta) > 0\}$ is the same for all $\theta \in \mathbb{R}$.

(A2) $l(\theta, x) = \log f(x|\theta)$ is thrice differentiable w.r.t. θ in the neighbour - hood $(\theta_0 - \delta, \theta_0 + \delta)$ of θ_0. If l', l'' and l''' be the 1st, 2nd and 3rd order derivatives then $E_{\theta_0}(l'(\theta_0))$ and $E_{\theta_0}(l''(\theta_0))$ are both finite, and

$$\sup_{\theta \in (\theta_0 - \delta, \theta_0 + \delta)} |l'''(\theta, x)| \le M(x) \text{ and } E_{\theta_0}(M(X_1)) < \infty.$$

(A3) Interchange of the order of expectation w.r.t. P_{θ_0} and differentiation at θ_0 are justified, so that

$$E_{\theta_0}(l'(\theta_0)) = 0, \quad E_{\theta_0}(l''(\theta_0)) = -E_{\theta_0}(l'(\theta_0))^2 = -I(\theta_0) \text{ (say)}$$

(A4) $0 < I(\theta_0) < \infty$.

(A5) For any $\delta > 0$

$$P_{\theta_0}\left[\text{For some } \epsilon > 0, \sup_{|\theta-\theta_0|>\delta} \frac{1}{n}\left[l_n(\theta) - l_n(\theta_0)\right] < -\epsilon \text{ for all sufficiently large } n\right] = 1.$$

(A6) $\theta \sim \pi(\theta)$ where, $\pi(\theta)$ is continuous and positive at θ_0.

Let $\hat{\theta}_n$ be a strongly consistent sequence of soln. of likelihood equation.

Let $\pi_n^*\left(t \mid x_1, \ldots, x_n\right)$ denote the posterior density of $t = \sqrt{n}\left(\theta - \hat{\theta}_n\right)$.

Then,

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \underbrace{\pi^*\left(t \mid x_1, \ldots, x_n\right)}_{\text{random quantity that's why prob. comes into play}} - \underbrace{\frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)}}_{\text{density of } N\left(0, \frac{1}{I(\theta_0)}\right)} \right| dt = 0$$

with prob. 1 under P_{θ_0}.

The evaluation is done considering θ_0 to be the true value.

Asymptotic Normality of Posterior distribution:-

Let $\tilde{\theta}_n$ be the posterior mode. $\underset{\sim}{X}_n = (X_1, \ldots, X_n)$

$\pi(\theta | \underset{\sim}{X}_n)$: posterior density

Under suitable regularity conditions, a Taylor series expansion of $\log \pi(\theta | \underset{\sim}{X}_n)$ at $\hat{\theta}_n$ gives —

$$\log \pi(\theta | \underset{\sim}{X}_n) = \underbrace{\log \pi(\tilde{\theta}_n | \underset{\sim}{X}_n)}_{\text{free of } \theta} + (\theta - \hat{\theta}_n) \frac{\partial}{\partial \theta} \log \pi(\theta | \underset{\sim}{X}_n) \Big|_{\tilde{\theta}_n} - \frac{1}{2} (\theta - \tilde{\theta}_n) \tilde{I}_n (\theta - \tilde{\theta}_n) + \ldots\ldots$$

where, $\tilde{I}_n$ is a $p \times p$ matrix defined as follows:

$$\log \pi(\theta | \underset{\sim}{X}_n) \simeq \log \pi(\hat{\theta}_n | \underset{\sim}{X}_n) - \frac{1}{2} (\theta - \hat{\theta}_n)' I_n (\theta - \tilde{\theta}_n)$$

$$\pi(\theta | \underset{\sim}{X}_n) \propto \exp\left[-\frac{1}{2} (\theta - \tilde{\theta}_n)' \tilde{I}_n (\theta - \tilde{\theta}_n) \right] \quad \text{(approximately)}$$

a $N_p(\tilde{\theta}_n, \tilde{I}_n^{-1})$ density for θ.

$$\tilde{I}_n = \left(-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\underset{\sim}{X}_n | \theta) \right)_{\tilde{\theta}_n = \text{MLE}}$$ is the observed Fisher Information matrix (generalised observed Fisher Information matrix).

Result: Under suitable regularity assumptions, for large n, the posterior distn. of θ can be approximated by any one of the normal distn. $N_p(\tilde{\theta}_n, \tilde{I}_n^{-1})$ or $N_p(\tilde{\theta}_n, I^{-1}(\tilde{\theta}_n))$. In particular, under suitable regularity conditions, the posterior distn. of $\hat{I}_n^{1/2} (\theta - \tilde{\theta}_n)$ gives $\underset{\sim}{X}_n$ converges to $N_p(0, I)$; where, I = Identity matrix of order p.

— We now formally state a theorem giving a set of regularity conditions under which asymptotic normality of posterior distn. holds.

— Let $X_1, \ldots, X_n$ be iid obsn with common distn. P_θ possessing a density $f(x|\theta)$ where $\theta \in (H)$, an open subset of $\mathbb{R}$. We fix some $\theta_0 \in (H)$ which may be regarded as the "true" value of the parameter as the probability statements are all made under θ_0.

Let $l(\theta, x) = \log f(x|\theta)$

$$L_n(\theta) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} l(\theta, x_i)$$: log-likelihood function.

For a function, $h(\cdot)$, let $h^{(i)}$ denotes the i^{th} derivative of h.

This is also known as Bernestein - Von-Misses theorem or Bayesian Central Limit theorem.

The assumptions are as follows:

(A1) The set $\{x : f(x|\theta) > 0\}$ is the same for all $\theta \in \Theta$

(A2) $l(\theta, x)$ is thrice differentiable w.r.t. θ in a neighbourhood $(\theta_0 - \delta, \theta_0 + \delta)$ of θ_0. the expectations $E_{\theta_0} l^{(1)}(\theta_0, X_1)$ and $E_{\theta_0} l^{(2)}(\theta_0, X_1)$ are both finite and

$$\sup_{\theta \in (\theta_0 - \delta, \theta_0 + \delta)} |l^{(3)}(\theta, x)| \le M(x) \text{ and } E_{\theta_0}(M(X_1)) < \infty.$$

(A3) Interchange of the order of expectation w.r.t. P_{θ_0} and differentiation at θ_0 are justified so that

$$E_{\theta_0} l^{(1)}(\theta_0, X_1) = 0,$$

$$E_{\theta_0} l^{(2)}(\theta_0, X_1) = -E_{\theta_0}\left[l^{(1)}(\theta_0, X_1)\right]^2 = -I(\theta_0),$$

where, $I(\theta_0)$ is the Fisher Information number per unit observation is positive and finite.

(A4) For any $\delta > 0$ with P_θ - probability one

$$\sup_{|\theta - \theta_0| > \delta} \frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right] < -\epsilon$$

for some $\epsilon > 0$ and for all sufficiently large n.

Remark:- We assume that there is a strongly consistent solution $\hat{\theta}_n$ to the likelihood equation $L_n^{(1)}(\theta) = 0$, i.e., $\exists$ a sequence of statistics $\hat{\theta}_n \ni$ with P_θ - probability one $\hat{\theta}_n$ satisfies the likelihood equation for sufficiently large n and $\hat{\theta}_n \to \theta_0$.

Theorem:- Suppose assumption (A1)-(A4) hold and $\hat{\theta}_n$ is a strongly consistent solution of the likelihood equation, then for any prior density $\pi(\theta)$ which is continuous and positive at θ_0,

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \pi_n^*(t \mid x_1, \ldots, x_n) - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt = 0 \quad \text{-----} (*)$$

with P_θ - probability one, where, $\pi_n^*(t \mid x_1, \ldots, x_n)$ is the posterior density of $t = \sqrt{n}(\theta - \hat{\theta}_n)$ given $x_1, \ldots, x_n$.

Also, under the same assumptions, (*) holds with $I(\theta_0)$ replaced by $\hat{I}_n = -\frac{1}{n} L_n^{(2)}(\hat{\theta}_n)$.

Proof:

Posterior density of θ given $x_1, \ldots, x_n$ is $\pi_n(\theta \mid x_1, \ldots, x_n)$ which equals to

$$\pi_n(\theta \mid x_1, \ldots, x_n) = \frac{\prod_{i=1}^n f(x_i \mid \theta)\, \pi(\theta)}{\int \prod_{i=1}^n f(x_i \mid \theta')\, \pi(\theta')\, d\theta'}$$

[Putting $\theta = \theta_n + \frac{t}{\sqrt{n}}$]

Posterior distn of $t = \sqrt{n}(\theta - \hat{\theta}_0)$ is

$$\pi_n^*(t \mid x_1, \ldots, x_n) = \frac{\pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \prod_{i=1}^n f\left(x_i \mid \hat{\theta}_n + \frac{t}{\sqrt{n}}\right)}{\int \prod_{i=1}^n f\left(x_i \mid \hat{\theta}_n + \frac{t'}{\sqrt{n}}\right) \pi\left(\hat{\theta}_n + \frac{t'}{\sqrt{n}}\right) dt'}$$

$$\therefore \pi_n^*(t \mid \underset{\sim}{x}_n) = \frac{\pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \left[\prod_{i=1}^n f\left(x_i + \hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \Big/ \prod_{i=1}^n f(x_i \mid \hat{\theta}_n)\right]}{\int \pi\left(\hat{\theta}_n + \frac{t'}{\sqrt{n}}\right) \prod_{i=1}^n \frac{f\left(x_i + \hat{\theta}_n + \frac{t'}{\sqrt{n}}\right)}{f(x_i \mid \hat{\theta}_n)}\, dt'}$$

$$= \frac{\pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\theta}_n + t/\sqrt{n}\right) - L_n(\hat{\theta}_n)\right]}{\int \underbrace{\pi\left(\hat{\theta}_n + \frac{t'}{\sqrt{n}}\right) \exp\left\{L_n\left(\hat{\theta}_n + \frac{t'}{\sqrt{n}}\right) - L_n(\hat{\theta}_n)\right\}}_{C_n} dt'}$$

$$= C_n^{-1}\, \pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left\{L_n\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - L_n(\hat{\theta}_n)\right\}$$

[$L_n(\theta) = \log \pi \prod f(x_i \mid \theta)$
$= \sum \log f(x_i \mid \theta)$

Need to show,

(1) $\int_{\mathbb{R}} \left| C_n^{-1}\, \pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left\{ Ln\left(\hat{\theta} + \frac{t}{\sqrt{n}}\right) - Ln(\hat{\theta}_n) \right\} - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt \longrightarrow 0$ a.s. $[P_{\theta_0}]$

(almost surely)

We first note that it is enough to show:

(2) $\int_{\mathbb{R}} \Big| \pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left\{ Ln\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - Ln(\hat{\theta}_n) \right\} - \pi(\theta_0)\, e^{-\frac{1}{2}t^2 I(\theta_0)} \Big| dt \longrightarrow 0$ a.s.

$g_n(t)$, say

To see this, note that (1) equals to

(3) $C_n^{-1} \left[\int_{\mathbb{R}} \left| \pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left\{ Ln\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - Ln(\hat{\theta}) \right\} - C_n \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt \right.$

and as (2) implies $C_n \to \pi(\theta_0)\sqrt{\frac{2\pi}{I(\theta_0)}}$, it is enough to show that the integral $\overset{(I)}{\wedge}$ in (3) goes to 0 a.s.

But $0 \le I \le I_1 + I_2$; where,

$$I_1 = \int_{\mathbb{R}} \left| \pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left\{ Ln\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - Ln(\hat{\theta}_n) \right\} - \pi(\theta_0)\, e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt$$

$$I_2 = \int_{\mathbb{R}} \left| \pi(\theta_0)\, e^{-\frac{1}{2}t^2 I(\theta_0)} - C_n \sqrt{\frac{I(\theta_0)}{2\pi}}\, e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt$$

Now $I_1 \to 0$ a.s. by (2) and

$$I_2 = \left| \pi(\theta_0) - C_n \sqrt{\frac{I(\theta_0)}{2\pi}} \right| \int_{\mathbb{R}} e^{-\frac{1}{2}t^2 I(\theta_0)} dt \longrightarrow 0$$

as $C_n \to \pi(\theta_0)\sqrt{\frac{2\pi}{I(\theta_0)}}$.

Now we are to prove (2).

$$\int_{\mathbb{R}} |g_n(t)|\,dt \to 0 \quad a.s. \ [P_{\theta_0}]$$

where $g_n(t) = \pi\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - L_n(\hat{\theta}_n)\right] - \pi(\theta_0)\, e^{-\frac{1}{2}t^2 I(\theta_0)}$.

Note: $\int_{\mathbb{R}} |g_n(t)|\,dt = \int_{A_1} |g_n(t)|\,dt + \int_{A_2} |g_n(t)|\,dt$, where

$A_1 = \{t : |t| < \delta_0\sqrt{n}\}$ and $A_2 = \{t : |t| > \delta_0\sqrt{n}\}$

δ_0 is approximately chosen small (+)ve number.

[We choose δ_0 so small that $2\delta_0 < \delta$, $|\theta - \theta_0| < 2\delta_0$
$\pi(\theta) < 2\pi(\theta_0)$
$\frac{\delta_0}{\sigma} E_{\theta_0} M(x_1) < \frac{\pi(\theta_0)}{2}$]

Result1:- $\int_{A_1} |g_n(t)|\,dt \to 0$ a.s. $[P_{\theta_0}]$; where, $A_1 = \{t : |t| < \delta_0\sqrt{n}\}$

Note that with P_θ - probability one, the following holds:

$$\left\{ \hat{\theta}_n \to \theta_0,\ -\frac{1}{n} L_n''(\theta_0) \to I(\theta_0),\ \frac{1}{n}\sum_{i=1}^{n} M(x_i) \to E_\theta\left(M(x_i)\right) \right\} \cdots (*)$$

We fix a sample sequence $(x_1, x_2, \ldots)$ for which (*) hold and prove our result for this sequence

$$L_n\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - L_n(\hat{\theta}_n) = \frac{t^2}{2n} L_n''(\hat{\theta}_n) + \frac{1}{6}\left(\frac{t}{\sqrt{n}}\right)^2 L_n^{(3)}(\theta_n')$$

$$= -\frac{1}{2}t^2 \hat{I}_n + R_n(t), \text{ say, } \theta_n' \text{ lying between } \hat{\theta} + \frac{t}{\sqrt{n}} \text{ and } \hat{\theta}_n.$$

and $\hat{I}_n = -\frac{1}{n} L_n^{(2)}(\hat{\theta}_n)$.

So, $\hat{I}_n = -\frac{1}{n} L_n''(\hat{\theta}_n) = \underbrace{-\frac{1}{n} L''(\theta_0)}_{\to I(\theta_0)} \underbrace{- \frac{1}{n}(\hat{\theta}_n - \theta_0) L_n^{(3)}(\theta_n'')}_{\to 0}$

where, θ_n'' lying between θ_0 and $\tilde{\theta}_n$.

$$\left[\text{As } \left|\frac{1}{n} L_n^{(3)}(\theta_n'')\right| \le \frac{1}{n}\sum_{i=1}^{n} M(X_i) \;\forall n \;(\text{for sufficiently large } n)\right.$$
$$\left.\longrightarrow E_\theta(M(X_1)) < \infty \quad \text{by (A3)}\right]$$

$\therefore \hat{I}_n \to I(\theta_0)$

For each fixed t,

$$|R_n(t)| = \frac{|t|^3}{6\sqrt{n}} \left|\frac{1}{n} L_n^{(3)}(\theta_n')\right|$$
$$\le \frac{|t|^3}{6\sqrt{n}} \times \frac{1}{n}\sum_{i=1}^{n} M(X_i) \quad \forall n$$
$$\to 0$$

Thus for each fixed t, $g_n(t) \to 0$.

For fix $t \in A$,

$$|R_n(t)| = \left(\frac{|t|}{\sqrt{n}}\right)^3 \times \frac{1}{6}\left|L_n^{(3)}(\theta_n')\right|$$
$$\le \delta_0 \cdot \frac{t^2}{6}\left|\frac{1}{n} L_n^{(3)}(\theta_n')\right|$$
$$\le \delta_0 \times \frac{t^2}{6} \times \frac{1}{n}\sum_{i=1}^{n} M(X_i) \quad \forall n.$$

$$\left[\; |\hat{\theta}_n + \tfrac{t}{\sqrt{n}} - \theta_0| < 2\delta_0, \; |\hat{\theta}_n - \theta_0| < \delta_0 \;\forall n \;\right]$$
$$\Rightarrow |\theta_n' - \theta_0| < 2\delta_0 < \delta \;\forall n$$

Now, $\frac{\delta_0}{6} \times \frac{1}{n}\sum_{i=1}^{n} M(X_i) < \frac{1}{4}\hat{I}_n \;\forall n$

$$\left[\begin{array}{l}\text{since LHS} \to \frac{\delta_0}{6} E_{\theta_0} M(X_i) \\ \qquad < \frac{1}{8} I(\theta_0) \\ \text{RHS} \to \frac{1}{4} I(\theta_0)\end{array}\right.$$

For $t \in A$, $|R_n(t)| < \frac{1}{4} t^2 \hat{I}_n \;\forall n$

$$\therefore L_n\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - L_n(\hat{\theta}_n) \le \frac{1}{2} t^2 \hat{I}_n + \frac{1}{4} t^2 \hat{I}_n = -\frac{1}{4} t^2 \hat{I}_n.$$

$$\exp\left(L_n\left(\hat{\theta}_n + \frac{t}{\sqrt{n}}\right) - L_n(\hat{\theta}_n)\right) < e^{-\frac{1}{4} t^2 \hat{I}_n} < e^{-\frac{1}{8} t^2 I(\theta_0)}$$
$$\left[\because \hat{I}_n > \tfrac{1}{2} I(\theta_0) \;\forall n\right]$$

Then

$$|g_n(t)| \le 2\pi(\theta_0) e^{-\frac{1}{8} t^2 I(\theta_0)} + \pi(\theta_0) e^{-\frac{1}{2} t^2 I(\theta_0)}$$

We have proved that $\int_{A_1} |g_n(t)|\, dt \to 0$ a.s. $[P_{\theta_0}]$

Now we will prove that $\int_{A_2} |g_n(t)|\,dt \to 0$ a.s. $[P_{\theta_0}]$.

Note, $\int_{A_2} |g_n(t)|\,dt \le \int_{A_2} \underbrace{\pi\left(\hat\theta_n + \frac{t}{\sqrt n}\right) \exp\left\{L_n\left(\hat\theta_n + \frac{t}{\sqrt n}\right) - L_n(\hat\theta_n)\right\}}_{< e^{-\frac{n\epsilon}{2}} + \int_{A_2} \pi(\theta_0)\, e^{-\frac12 t^2 I(\theta_0)}\, dt} dt$

$\to 0$ (normal tail) ---- (3)

For $t \in A_2$,

$$\frac1n\left\{L_n\left(\hat\theta_n + \frac{t}{\sqrt n}\right) - L_n(\hat\theta_n)\right\} = \frac1n\left\{L_n\left(\hat\theta_n + \frac{t}{\sqrt n}\right) - L_n(\theta_0)\right\} + \frac1n\left\{L_n(\theta_0) - L_n(\hat\theta_n)\right\}$$

$\left[\text{For } t \in A_2,\ \left|\hat\theta_n + \frac{t}{\sqrt n} - \theta_0\right| > \frac{\delta_0}{2}\ \forall\, n\right]$

$$\le \sup_{|\theta-\theta_0|>\frac{\delta_0}{2}} \frac1n\left[L_n(\theta) - L_n(\theta_0)\right] + \frac1n\left\{\frac12(\theta_0 - \hat\theta_n)^2 L_n''(\theta_n) + \frac{(\theta_0 - \hat\theta_n)^3}{6} L_n^{(3)}(\theta_n''')\right\},$$

where, θ_n''' lies between θ_0 and $\hat\theta_n$.

$$< -\frac{\epsilon}{2} \quad \forall\, n$$

By assumption (A4), with P_θ-probability one,

$$\sup_{|\theta-\theta_0|>\frac{\delta_0}{2}} \frac1n\left[L_n(\theta) - L_n(\theta_0)\right] < -6 \text{ for some } \epsilon > 0\ \forall\, n \quad \text{------ (**)}$$

Suppose the sample sequence $(X_1, X_2, \ldots)$ satisfies (**) in addition to (*).

Then $\forall\, t \in A_2$, $\frac1n\left\{L_n\left(\hat\theta_n + \frac{t}{\sqrt n}\right) - L_n(\hat\theta_n)\right\} < -\epsilon/2\ \forall\, n.$

Now from (3),

$$\int_{A_2} |g_n(t)|\,dt \le \int_{A_2} \pi\left(\hat\theta_n + t/\sqrt n\right) e^{-n\epsilon/2}\, dt + \int_{A_2} \pi(\theta_0)\, e^{-\frac12 t^2 I(\theta_0)}\, dt$$

$$\le \underbrace{e^{-n\epsilon/2}\sqrt n \int_{\mathbb R} \pi(\theta)\, d\theta}_{\to 0} + \underbrace{\int_{A_2} \pi(\theta_0)\, e^{-\frac12 t^2 I(\theta_0)}\, dt}_{\to 0 \text{ (normal tail)}}$$

Proof of 2nd Part:-

To show $\int_{\mathbb{R}} \left| \pi^*(t \mid x_1, \ldots, x_n) - \frac{\sqrt{\hat{I}_n}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 \hat{I}_n} \right| dt \to 0$ a.s.

$$\leq \int_{\mathbb{R}} \left| \pi^*(t \mid x_1, \ldots, x_n) - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt$$

$$+ \int_{\mathbb{R}} \underbrace{\left| \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} - \frac{\sqrt{\hat{I}_n}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 \hat{I}_n} \right| dt}_{\to 0 \text{ by DCT}}$$

($\to 0$, 1st part)

Here we completed the proof of BVM Theorem.

The following theorem states that in the regular case, with the large sample, a Bayes estimate is approximately same as the MLE $\hat{\theta}_n$.

If we consider the squared error loss, the Bayes estimate for θ is given by the posterior mean

$$\theta_n^* = \int \theta \pi_n(\theta \mid x_1, \ldots, x_n) d\theta$$

- **Theorem:-** In addition to the assumption of the previous theorem, assume that $\int |\theta| \pi(\theta) d\theta < \infty$.
Then $\sqrt{n}(\theta_n^* - \hat{\theta}_n) \to 0$ with probability one.

Proof:- Proceeding as in the proof of the previous theorem and using the assumption of finite mean of $\pi(\cdot)$ we can prove

$$\int_{\mathbb{R}} |t| \left| \pi_n^*(t \mid x_1, \ldots, x_n) - \sqrt{\frac{I(\theta_0)}{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} \right| dt \to 0 \text{ a.s.}$$ [CHECK]

This implies $\int t \pi_n^*(t \mid x_1, \ldots, x_n) dt \to \int t \sqrt{\frac{I(\theta_0)}{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)} dt$

Now, $\theta_n^* = E(\theta \mid x_1, \ldots, x_n) = 0$ a.s.

$= E\left(\hat{\theta}_n + t/\sqrt{n} \mid x_1, \ldots, x_n\right) = \hat{\theta}_n + \frac{1}{\sqrt{n}} \int t \pi_n^*(t \mid x_1, \ldots, x_n)$

$\therefore \sqrt{n}(\theta_n^* - \hat{\theta}_n) = \int t \pi_n^*(t \mid x_1, \ldots, x_n) dt \to 0$ a.s.

Remark:- $\sqrt{n}(\theta_n^* - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$ [Sampling Theory Result].

Suppose $X \sim P$ and $P \in \mathcal{P}$

Laplace has given the following idea of consistency:

If $X_1, X_2, \ldots$ iid Bernoulli (θ) and $\pi(\theta)$ is prior density that is continuous and (+ve) on $(0,1)$, then the posterior is consistent at all $\theta_0 \in (0,1)$.

An elementary proof of Laplace's result for a Beta Prior is as follows:

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\theta \mid X_1, \ldots, X_n \sim \text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha,\ n - \sum_{i=1}^{n} x_i + \beta\right)$$

$$E\left(\theta \mid X_1, \ldots, X_n\right) = \frac{\sum_{i=1}^{n} x_i + \alpha}{\alpha+\beta+n} \longrightarrow \theta_0 \text{ as } n \to \infty.$$

$$\text{Var}\left(\theta \mid X_1, \ldots, X_n\right) = \frac{\left(\sum_1^n x_i + \alpha\right)\left(n - \sum_1^n x_i + \beta\right)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \longrightarrow 0 \text{ as } n \to \infty$$

In most unusual examples, posterior distns are consistent. We consider now, the interesting situations when

$$(H) = \{\theta_1, \ldots, \theta_k\}, \text{ a finite set.}$$

Theorem:- Let $(H) = \{\theta_1, \ldots, \theta_k\}$ and $X_1, X_2, \ldots$ be iid observations with common density $f(x|\theta)$, $\theta \in (H)$.

Consider a prior $\{p_1, \ldots, p_k\}$ where $p_i > 0 \ \forall\, i = 1(1)k$ and $\sum_{i=1}^{k} p_i = 1$ and $p_i = \text{Pr}_0(\theta = \theta_i)$.

Suppose $\theta_t \in (H)$ is the true value of θ and θ_i's are distinguishable $\ni \forall\ i \neq t$

$$\int f(x|\theta_t) \log\left(\frac{f(x|\theta_t)}{f(x|\theta_i)}\right) dx > 0.$$

Then a.s.

$$\lim_{n\to\infty} p\left(\theta_t \mid x_1, \ldots, x_n\right) = 1 \text{ and}$$

$$\lim_{n\to\infty} p\left(\theta_i \mid x_1, \ldots, x_n\right) = 0, \quad i \neq t.$$

Define, $I(\theta) = E\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2$

If $I(\theta) \simeq 0 \Rightarrow \frac{\partial}{\partial\theta}\log f(x,\theta)$ is small.

If $I(\theta)$ is large, $\Rightarrow \frac{\partial}{\partial\theta}\log f(x,\theta)$ is large.

Hence when $I(\theta)$ is small, the rate of change of likelihood w.r.t. θ remains constant.

Proof:-

$$p(\theta_i|x_1,\ldots,x_n) = \frac{p_i f(x_1,\ldots,x_n|\theta_i)}{\sum_{j=1}^{K} p_j f(x_1,\ldots,x_n|\theta_j)}$$

$$= \frac{p_i \prod_{n=1}^{n} f(x_n|\theta_i)}{\sum_{j=1}^{K} p_j \prod_{n=1}^{n} f(x_n|\theta_j)}$$

To show $p(\theta_i|x_1,\ldots,x_n) \longrightarrow 1$ if $\theta_i = \theta_t$
$\longrightarrow 0$ if $\theta_i \neq \theta_t$

$$\therefore p(\theta_i|x_1,\ldots,x_n) = \frac{p_i \prod_{n=1}^{n} \frac{f(x_n|\theta_i)}{f(x_n|\theta_t)}}{\sum_{j=1}^{K} p_j \prod_{n=1}^{n} \frac{f(x_n|\theta_j)}{f(x_n|\theta_t)}}$$

$$= \frac{p_i e^{nS_i}}{\sum_{j=1}^{K} p_j e^{nS_j}}, \text{ where } S_i = \frac{1}{n}\sum_{n=1}^{n} \log\frac{f(x_n|\theta_i)}{f(x_n|\theta_t)}$$

$$= \frac{p_i e^{nS_i}}{p_t + \sum_{1\le j\neq t\le k} p_j e^{nS_j}}$$

The rest of the proof is skipped.

Doob's Theorem:- Suppose (Θ) and $\mathcal{X}$ are both complete separable metric spaces (or, polish spaces) endowed with their respective Borel σ-fields $B(\Theta)$ and $\mathcal{A}$ and let $\theta \to P_\theta$ be one to one.

Let, Π be a prior and $\{\Pi(\cdot|x_1,\ldots,x_n)\}$ be a posterior. Then $\exists\ (\Theta)_0 \subset (\Theta)$ with $\Pi((\Theta)_0) = 1$ s.t. $\{\Pi(\cdot|x_1,\ldots,x_n)\}_{n\geq 1}$ is consistent at every $\theta \in (\Theta)_0$

$$P_{\theta_0}^{\infty}\left(\sqrt{n}(\theta-\theta_0)\right)_{\underset{\sim}{X}} \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

Bayesian Information Criterion (BIC):-

Consider a model with likelihood $f(\underset{\sim}{x}|\underset{\sim}{\theta})$ and prior $\Pi(\underset{\sim}{\theta})$. Then the integrated likelihood

$$m(\underset{\sim}{x}) = \int q(\theta)\, e^{nh(\theta)}\, d\theta,$$

where, $q(\theta) = \Pi(\underset{\sim}{\theta})$ and $nh(\underset{\sim}{\theta}) = \log f(\underset{\sim}{x}|\underset{\sim}{\theta})$.

The Laplace approximation formula can be used to find an approximation to $m(\underset{\sim}{x})$.

Schwarz (1978) proposed a criterion, known as the BIC based on this approximation, ignoring terms that stay bounded as the sample size $n \to \infty$.

The criterion is given by,

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

Note that this approximation is free from the choice of prior.

$$\hat{m}(x) = \frac{e^{\log f(\underset{\sim}{x}|\hat{\theta})}\, \Pi(\hat{\theta})\, (\sqrt{2\pi})^p}{\sqrt{|-n^p C(\hat{\theta})|}}$$

$$\log \hat{m}(\underset{\sim}{x}) = \log f(\underset{\sim}{x}|\hat{\theta}) - \frac{p}{2}\log n.$$

BAYESIAN HYPOTHESIS TESTING & MODEL SELECTION

On the irreconcilability of the p-value and Bayesian measures of evidence – an example.

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$

To test, $H_0: \theta = \theta_0$ Vs. $H_1: \theta \neq \theta_0$

We reject in frequentist inference if $T(\underset{\sim}{X}) = |\sqrt{n}(\bar{X} - \theta_0)|$ is

p-value $= P_{\theta_0}\left[|\sqrt{n}(\bar{X} - \theta_0)| > t\right]$, where, t is the observed value of $T(\underset{\sim}{X}) = |\sqrt{n}(\bar{x} - \theta_0)|$

i.e., $t = |\sqrt{n}(\bar{x} - \theta_0)|$

Assume, $P(H_0) = P(H_1) = 1/2$.

Consider a prior degenerate at θ_0 and H_0 and $N(\mu, \tau^2)$ prior for θ under H_1.

$$\pi(\theta) = \frac{1}{2} I(\theta = \theta_0) + \frac{1}{2} N(\mu, \tau^2) I(\theta \neq \theta_0)$$

B_{01} = Bayes Factor of H_0 to H_1

$$= \frac{f(\underset{\sim}{x} | \theta_0)}{\int f(\underset{\sim}{x} | \theta) \pi(\theta) d\theta}$$

Taking $\mu = \theta_0$, $\tau = 1$, we have

$$B_{01} = \sqrt{1+n} \exp\left\{-\frac{1}{2} \cdot \frac{t^2}{1+\frac{1}{n}}\right\}$$

Verify:

$$\sqrt{1+\frac{1}{\rho^2}} \exp\left[-\frac{1}{2}\left\{\frac{(t-\rho\eta)^2}{1+\rho^2} - \eta^2\right\}\right]$$

where, $\rho = \frac{1}{\sqrt{n}\tau}$, $\eta = \frac{(\theta_0 - \mu)}{\tau}$

Take $n = 50$, $t = 1.96$

p-value $= 0.05$, $B_{01} = 1.08$.

$P(H_0 | \underset{\sim}{x}) = 0.52$.

On the comparison of p-values and posterior probabilities:

$X_1, \ldots, X_n | \theta \overset{iid}{\sim} N(\theta, 1)$

$H_0: \theta = \theta_0$ Vs. $H_1: \theta \neq \theta_0$

Notation:- B = Bayes factor of H_0 relative to H_1

$\rho = P(H_0 | \underset{\sim}{x})$

n = sample size

		1		5		10		20		50		100	
t	α	B	P	B	P	B	P	B	P	B	P	B	P
1.645	0.10	0.72	0.42	0.79	0.44	0.89	0.47	1.27	0.56	1.86	0.66	-	-
1.960	0.05	0.54	0.35	0.49	0.33	0.59	0.37	1.08	0.52	1.08	0.52	-	-
2.576	0.01	0.27	0.21	0.15	0.13	0.16	0.14	0.28	0.22	0.28	0.22	-	-
3.291	0.001	0.10	0.09	0.03	0.03	0.02	0.02	0.03	0.03	0.03	0.03	-	-

Remarks:-

1. To a Bayesian the posterior probability of H_0 summarises the evidence against H_0. In many common cases of testing p-value is smaller that the posterior prob. of H_0. by an order of magnitude. It may be noted that p-value ignores the likelihood of the data under the alternative and takes into account not only the observed deviation of the data from null hypothesis. as measured by test statistic but also more significant deviation.

2. One may try to see what happens when one changes the prior density of θ under H_1. We will see that the difference between p-value and posterior probabilities persists even when we consider lower bounds on B and P over large class of priors under the alternative.

Note: We note that the differences between p-values and the corresponding Bayesian measures of evidence remain irreconcilable even when the lower bounds on such measures are considered. In other words, the least possible Bayes factors and posterior prob. of H_0 are substantially larger than the corresponding p-values.

Suppose we are comparing K competing models: $M_1, M_2, \ldots, M_K$. for data X such that $\{M_i : X \text{ has density } f_i(\underset{\sim}{x} | \theta_i)$ for $i = 1(1)K$. Here θ_i are the unknown model parameters. We will use hypothesis testing and model selection almost interchangably in this discussion.

Eg: The hypothesis testing problem. Suppose $X_1, \ldots, X_n$ have joint densities $f(\underset{\sim}{x} | \theta)$, where $\theta \in (H)$, (H) being the model space.

Example 1: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$

Suppose we want to test $H_0 : \theta \leq \theta_0$ Vs. $H_1 : \theta > \theta_0$.
H_0 and H_1 induce a partition of the parameter space as
$(H) = (H)_0 \cup (H)_1$, where $(H)_0 = \{\theta \leq \theta_0\}$ and $(H)_1 = \{\theta > \theta_0\}$.
This can be framed also as a model selection problem where we want to choose between:

$M_0 : X_1, \ldots, X_n | \theta \overset{iid}{\sim} f(\underset{\sim}{x} | \theta)$ with $\theta \in (H)_0$
$M_1 : X_1, \ldots, X_n | \theta \overset{iid}{\sim} f(\underset{\sim}{x} | \theta)$ with $\theta \in (H)_1$

Example 2: Regression Models:

Y is the "dependent" variable and $x_1, x_2, \ldots, x_p$ are the "potential" regressors. If we use all the regressors the "full" linear regression model is:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \quad \text{where } \varepsilon_i \overset{iid}{\sim} N(0,1)$$

One possible class of models is the nested class.

$M_0 : \beta_0 \in \mathbb{R}$, $\beta_i = 0 \quad \forall i = 1(1)p$
$M_1 : (\beta_0, \beta_1) \in \mathbb{R}^2$, $\beta_i = 0 \ \forall i = 2(1)p$
$M_2 : (\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$, $\beta_i = 0 \ \forall i = 3(1)p$
$M_p : (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ (the full model)

$M_0 \subset M_1 \subset \cdots \subset M_p$.

Example 3: $X \sim f(x|\theta)$, $\theta \in \mathbb{R}$

M_0 : $f(x|\theta) = \text{Cauchy}(\theta, 1)$

M_1 : $f(x|\theta) = \text{Normal}(\theta, 1)$.

Suppose given models : $M, M_2, \ldots, M_K$. We have prior distn. $\pi_i(\theta_i)$, $i = 1(1)K$ for the unknown parameters. Define the marginal distn. or predictive distn. of X given by

$$m_i(\underset{\sim}{x}) = \int f_i(\underset{\sim}{x}|\theta_i)\pi_i(\theta_i)\, d\theta_i \quad , \quad i = 1(1)K.$$

The Bayes factor of M_j to M_i is given by

$$B_{ji}(\underset{\sim}{x}) = \frac{m_j(\underset{\sim}{x})}{m_i(\underset{\sim}{x})} = \frac{\int f_j(\underset{\sim}{x}|\theta_j)\pi_j(\theta_j)\, d\theta_j}{\int f_i(\underset{\sim}{x}|\theta_i)\pi_i(\theta_i)\, d\theta_i}$$

The Bayes factor is often interpreted as the odds provided by the data for M_j in comparison to M_i. Thus $B_{ji} = 10$ would suggest that the data favour M_j over M_i at odds of 10 to one.

Alternatively, B_{ji} may also be thought of weighted likelihood ratio of M_j to M_i, the priors being the weight function.

If prior probabilities $P(M_j)$, $j = 1(1)K$ of the models are available one can compute the posterior probability of the models from the Bayes factors. It is easy to see that the posterior prob. of M_i given $\underset{\sim}{x}$ is

$$P(M_i|\underset{\sim}{x}) = \frac{P(M_i)\, m_i(\underset{\sim}{x})}{\sum_{j=1}^{K} P(M_j)\, m_j(\underset{\sim}{x})} = \left[\sum_{j=1}^{K} \frac{P(M_j)}{P(M_i)} B_{ji}\right]^{-1} \quad \cdots (*)$$

A particularly common choice (although not necessarily the best option in complex problem) is

$$P(M_i) = \frac{1}{k} \quad \text{for} \quad i = 1(1)k.$$

In this case $P(M_i|\underset{\sim}{x}) = \dfrac{m_i(\underset{\sim}{x})}{\sum_{j=1}^{k} m_j(\underset{\sim}{x})} = \bar{m}_i(\underset{\sim}{x})$.

In scientific reporting it is common to provide $\bar{m}_i(\underset{\sim}{x})$ rather than $P(M_j|\underset{\sim}{x})$ since the prior probabilities of models can be a contentious matter and anyone can use the $\bar{m}_i(\underset{\sim}{x})$ to determine their personal posterior probabilities via (*), noting that $B_{ji} = \dfrac{\bar{m}_j(\underset{\sim}{x})}{\bar{m}_i(\underset{\sim}{x})}$

<u>Simple Case</u>: $K = 2$

$$B_{10} = \frac{m_1(\underset{\sim}{x})}{m_0(\underset{\sim}{x})}$$

$$P(M_0) = \pi_0 \quad P(M_1) = 1 - \pi_0$$

$$P(M_1|\underset{\sim}{x}) = \frac{P(M_1)\, m_1(\underset{\sim}{x})}{\sum_{j=0}^{k} P(M_j)\, m_j(\underset{\sim}{x})}$$

$$\frac{P(M_1|\underset{\sim}{x})}{P(M_0|\underset{\sim}{x})} = B_{10}\,\frac{P(M_1)}{P(M_0)}$$

$$P(M_0|\underset{\sim}{x}) = \frac{P(M_0)\, m_0(\underset{\sim}{x})}{\sum_{j=0}^{k} P(M_j)\, m_j(\underset{\sim}{x})}$$

i.e., $B_{10} = \dfrac{\text{posterior odds}}{\text{prior odds}}$

$$\frac{c_1 \int f_i(\underset{\sim}{x}|\theta_i)\, \pi_i^N(\theta_i)\, d\theta_i}{c_2 \int f_j(\underset{\sim}{x}|\theta_j)\, \pi_j^N(\theta_j)\, d\theta_j}$$

measure of evidence changes for same data, So, using non-informative prior doesn't justify the problem.

Motivation for Bayesian approach to model selection :—

1. Bayes factors and posterior model probability are easy to understand. The interpretation of Bayes factors are easily understandable by non-statisticians.
2. Bayesian model selection is consistent under mild condition in finite/fixed dimensional problem.
3. Bayesian procedures naturally penalize model complexity and they do need an explicit introduction of the penalty term. In this way, simpler models are preferred to complex models.
4. The Bayesian approach to model selection is conceptually the same, regardless of the no. of models under consideration.

Example: $H_0 : \theta \in \textcircled{H}_0$ Vs. $H_1 : \theta \in \textcircled{H} - \textcircled{H}_0 = \textcircled{H}_1$.

Case 1: $\textcircled{H}_0$ and $\textcircled{H}_1$ are of same "dimension"

(viz. $H_0 : \theta \le \theta_0$ Vs. $H_1 : \theta > \theta_0$ where $X_1, \ldots, X_n | \theta \overset{iid}{\sim} N(\theta, 1)$, $\theta \sim N(\mu, \tau^2)$

We will calculate $P(H_0 | \underset{\sim}{X})$ and $P(H_1 | \underset{\sim}{X})$ and check which one is larger.

$$\pi(\theta \le \theta_0) = \int_{\theta \le \theta_0} \pi(\theta) d\theta = \pi_0 \quad \text{and} \quad \pi(\theta > \theta_0) = 1 - \pi_0 .$$

$$\pi(\theta) = P(H_0) P(\theta | H_0) + P(H_1) P(\theta | H_1)$$

$$= \theta_0 \cdot \frac{N(\mu, \tau^2)}{\pi_0} + (1 - \pi_0) \frac{N(\mu, \tau^2)}{(1 - \pi_0)}$$

$$= \pi_0 \pi_1(\theta) I(\theta \in \textcircled{H}_0) + (1 - \pi_0) \pi_2(\theta) I(\theta \in \textcircled{H}_1), \text{ where}$$

$$\pi_1(\theta) = \frac{\exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right)}{\sqrt{2\pi}\, \tau \pi_0},$$

Case 2: $\textcircled{H}_0$ and $\textcircled{H}_1$ are not of same dimension, viz.

$H_0 : \theta = \theta_0$ Vs. $H_1 : \theta \neq \theta_0$, where,

$X_1, \ldots, X_n | \theta \overset{iid}{\sim} N(\theta, 1)$

$$\pi(\theta) = \pi_0 I(\theta = \theta_0) + (1 - \pi_0) \pi_1(\theta) I(\theta \neq \theta_0)$$

Motiration behind objective Bayerian model selection; —

There has been a long debate in the Bayesian community about the roles of subjective and objective Bayesian analysis. Subjective Bayesian analysis is an attractive proposition but few would disagree that it is frequently not a realistic possibility because it is sometimes simply not possible to obtain the extensive needed elicitation from subjective experts.

In model selection, this last argument is particularly compelling because we often initially entertain a wide variety of models and careful subjective specification of prior dists for all the parameters of the model is essentially impossible. Four methods for developing default Bayesian model selection have come in modern research. There are the conventional prior approach, the BIC approach, Intrinsic Bayes factor (IBF) approach, the fractional Bayes Factor approach.

Conventional prior approach:- $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$; μ, σ^2 unknown.

to test $H_0: \mu = 0$ Vs. $H_1: \mu \neq 0$

Jeffry's suggestion: Take prior $g_0(\sigma) = \frac{1}{\sigma}$ for σ under H_0, Under H_1, take the same prior for σ and take the conventional prior for μ and σ as $g_1(\mu|\sigma) = \frac{1}{\sigma} g_2(\mu|\sigma)$, where $g_2(\cdot)$ is a pdf. Then under H_1 the prior is

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

An initial choice of $g_2(\cdot)$ is $N(0, \tau^2)$, one can take $c = 1$.
Jeffrey noted that for reasonable prior specification $BF_{01} \to 0$ if $\bar{x} \to \infty$ and $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ is bounded.
He gave an argument which implies that unless g_1 has no finite moments this will not happen. In particular,

g_2 = Normal $BF_{01} \not\to 0$ under above criterion.

Jeffry suggested to use

g_2 = cauchy

So, Jeffry's final recommendation are

$g_0(\sigma) = \frac{1}{\sigma}$ under H_0

$g_1(\mu,\sigma) = \frac{1}{\sigma} g_2(\mu/\sigma) = \frac{1}{\sigma} \cdot \frac{1}{\sigma\pi\left(1+\frac{\mu^2}{\sigma^2}\right)}$ under H_1.

Jeffrey's test in normal problems:

$X_1, \ldots, X_n \overset{iid}{\sim} N(0, \sigma^2)$

$H_0: \mu = 0$ Vs. $H_1: \mu \neq 0$

Prior suggested: $g_0(\sigma) = \frac{1}{\sigma}$ under H_0

$g_1(\mu,\sigma) = \frac{1}{\sigma} \cdot \frac{1}{\sigma\pi\left(1+\frac{\mu^2}{\sigma^2}\right)}$ under H_1.

$$BF_{01} = \frac{\int f(x_1, \ldots, x_n | 0, \sigma^2) g_0(\sigma) d\sigma}{\int\int f(x_1, \ldots, x_n | \mu, \sigma^2) g_1(\mu, \sigma) d\mu d\sigma}$$

There is no closed form for the expression for the denominator of BF_{01}. To deal with it one may note that

$$g_1(\mu|\sigma^2) = \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\lambda}{2\sigma^2}\mu^2\right) \frac{1}{\sqrt{2\pi}} \lambda^{1/2-1} e^{-\lambda/2} d\lambda.$$

i.e., $\mu|\lambda \sim N(0, \sigma^2/\lambda)$, $\lambda \sim$ Gamma$(1/2, 1/2)$.

To calculate the dinominator of BF_{01}, one can integrate over μ and σ in closed form using the mixture representation. Ultimately, one has a one-dimensional integral over λ left which is easier to handle.

A curious Example:- Einstein's theory of gravitation predicts deflection of light due to gravitation and specifies the amount of deflection. He predicted that light of stars would deflect stars, under gravitational pull of the sun on the nearby stars, but the effect could be invisible only during a total solar eclipse when the deflection can be measured through the apparent change in star's position.

A famous experiment by a team lead by Eddington, led to acceptance of Einstein's theory. There are four observations, two collected in 1919 in Eddington's expedition and two more by other groups by in 1922 and 1929. These are $x_1 = 1.98$, $x_2 = 1.61$, $x_3 = 1.18$ and $x_4 = 2.24$.

Suppose they are iid $N(\mu, \sigma^2)$, both parameters unknown. Einstein's theory says μ should be 1.75.

We test $H_0: \mu = 1.75$ Vs. $H_1: \mu \neq 1.75$

If we use the conventional priors of Jeffrey's then BF_{01} turns out to be 2.98.

A common choice of prior for the normal linear model is the conjugate prior, called a g-prior in Zellner (1986). For a linear model $\mathcal{M}: y = X\beta + \epsilon$, $\epsilon \sim N_n(0, \sigma^2 I_n)$ where, σ^2 and $\beta = (\beta_1, \ldots, \beta_k)$ are unknown and X as $(n \times k)$ given design matrix of rank $k < n$, the g-prior density is given by

$$\pi(\sigma) = \frac{1}{\sigma}, \quad \pi(\beta | \sigma) \sim N_k\left(0, g\sigma^2 (X'X)^{-1}\right)$$

The key advantage of g-prior is that the marginal density, $m(y)$, is available in closed form and is given by,

$$m(y) = \frac{\Gamma(n/2)}{2\pi^{n/2}(1+g)^{k/2}} \left(y'y - \frac{g}{1+g} y'X(X'X)^{-1}X'y\right)^{-n/2}$$

Thus BF and posterior model probability for comparing any two linear models are available.

$$BF_{01}(\underset{\sim}{x}) = \frac{\int f_1(\underset{\sim}{x}|\theta_1)\pi_1(\theta_1)\,d\theta_1}{\int f_2(\underset{\sim}{x}|\theta_2)\pi_2(\theta_2)\,d\theta_2}$$

$$BF_{01}(\underset{\sim}{x}_2) = \frac{\int f_1(x_2|\theta_1)\pi(\theta_1|x_1)\,d\theta_1}{\int f_2(x_2|\theta_2)\pi(\theta_2|x_1)\,d\theta_2}$$

$\underset{\sim}{x} = (x_1, x_2)$ — x_1: updating prior, x_2: predicting

The Intrinsic Bayes Factor Approach: Comider two models M_0 and M_1 for data X with density $f_i(x|\theta_i)$ under M_i, θ_i being an unknown parameter of dimension p_i, $i=0,1$. Given prior specification $g_i(\theta_i)$ for parameter θ_i, the Bayes Factor of M_1 to M_0 is

$$B_{10} = \frac{m_1(\underset{\sim}{x})}{m_0(\underset{\sim}{x})} = \frac{\int f_1(x|\theta)g_1(\theta)\,d\theta}{\int f_0(x|\theta_0)g_0(\theta_0)\,d\theta} \quad \text{.....} (*)$$

When subjective specification of prior distn. is impossible, one would look for automatic method that uses standard non-informative priors that are typically improper.

If g_i's are improper, they are defined only upto arbitrary multiplicative constant c_i. Then $c_i g_i$ has as much validity as g_i. This implies that $\left(\frac{c_1}{c_0}\right)B_{10}$ has as much validity as B_{10}. Thus the Bayes factor is determined only upto arbitrary multiplicative constants.

Remark: It is a serious fact that the use of diffuse (flat) proper prior may not provide a good solution to the problem described about the use of proper prior. Truncation of non-informative priors leads to large penalty for the more complex model.

BAYESIAN COMPUTATION

Bayesian analysis requires computation of expectations and quantiles of probability distributions arising as posterior distributions. Sometimes modes of the posterior are also needed. If conjugate priors are not used, as in many cases, posterior distributions will not be standard distns. and hence the required Bayesian quantiles of interest can't be computed in closed form. So thus special techniques are needed for bayesian computation.

Example: Suppose $X \sim N(\theta, 1)$ and $\theta \sim$ cauchy (μ, τ) prior with μ, τ known. (say, $\mu = 0, \tau = 1$)

$$\pi(\theta|x) \propto \exp\left\{-\frac{(\theta-x)^2}{2}\right\}\left[\tau^2+(\theta-\mu)^2\right]^{-1}$$

$$E(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta \exp\left\{-\frac{(\theta-x)^2}{2}\right\}\left[\tau^2+(\theta-\mu)^2\right]^{-1} d\theta}{\int_{-\infty}^{\infty} \exp\left\{-\frac{(\theta-x)^2}{2}\right\}\left[\tau^2+(\theta-\mu)^2\right]^{-1} d\theta}$$

Easy to see that $E(\theta|x)$ does not come out in any analytical tractable form. It is clear that some form of approximation will be needed. One may try numerical integration techniques as the problem is low-dimensional.

Example: Suppose $X_1, \ldots, X_K$ are independent poisson counts with $X_i \sim$ Poisson (θ_i), θ_i's are a-priori considered related, and a joint multivariate normal prior distns as their logarithm is assumed. Specifically, let $\gamma_i = \log(\theta_i)$ and suppose $\underset{\sim}{\gamma} = (\gamma_1, \ldots, \gamma_K)$ has a distn $\underset{\sim}{\gamma} \sim N_K\left(\mu \underset{\sim}{1}, \tau^2\{(1-\rho) I_K + \rho \underset{\sim}{1}\underset{\sim}{1}'\}\right)$ where, $\underset{\sim}{1}$ is the K-vector with all element being 1 (μ, τ^2, ρ are known).

$$\pi\left(\underset{\sim}{\gamma}|\underset{\sim}{x}\right) \propto \exp\left\{-\sum_{i=1}^{K}\{e^{\gamma_i} - \gamma_i x_i\} - \frac{1}{2\tau^2}\left(\underset{\sim}{\gamma}-\mu\underset{\sim}{1}\right)' \{(1-\rho) I_K + \rho \underset{\sim}{1}\underset{\sim}{1}'\}^{-1}\left(\underset{\sim}{\gamma}-\mu\underset{\sim}{1}\right)\right\}$$

$$P(X|\theta) = \frac{e^{-\theta}\theta^x}{x!} = e^{-\theta} e^{x\log\theta}.$$

$$E(\theta_j | \underset{\sim}{x}) = E\left(\exp(\gamma_j) | \underset{\sim}{x}\right)$$

$$= \frac{\int_{\mathbb{R}^k} \exp(\gamma_j)\, g(\underset{\sim}{\gamma} | \underset{\sim}{x})\, d\underset{\sim}{\gamma}}{\int_{\mathbb{R}^k} g(\underset{\sim}{\gamma} | \underset{\sim}{x})\, d\underset{\sim}{\gamma}}$$

(Curse of dimensionality) The above is a ratio of two k-dimensional integrals and as k grows, the integrals become less easy to work with. Numerical integration fails to be an efficient option in that case.

The errors in approximation associated with the numerical methods increase as a power of the dimensions in numerical integration may not be a good option except in very very low dimensional problem.

The way out of these problems, in many cases is through other approximation techniques. One could try, in some cases, the laplace approximation, EM algorithm etc. But here we will stress on general purpose simulation based techniques.

Monte Carlo Sampling :- An alternative to numerical integration or analytical approximation to compute an integral is statistical sampling. This probabilistic technique is a familiar tool in statistical inference. To estimate a population mean or a population proportion, a natural approach is to gather a large sample from this population and to consider the corresponding sample mean or sample proportion. The laws of large numbers guarantee that the estimates thus obtained are good provided the sample size is large.

Ref: "Markov chain Monte carlo in Practice" by Spigelhalter et al.

Suppose f be a density (or pmf) and we want to estimate

$$E_f h(x) = \int_{\mathcal{X}} h(x) f(x) dx .$$

If iid observation, $X_1, X_2, \ldots$ can be generated from f, then

$$\overline{h}_m = \frac{1}{m} \sum_{i=1}^{m} h(x_i) \xrightarrow[a.s.]{P} E_f h(X) .$$

This provides a justification of using $\overline{h}_m$ as an approximation for $E_h f(x)$ for large m.

Example (Continued)

$$E(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta \, \phi(\theta - x) \left(\tau^2 + (\theta-\mu)^2\right)^{-1} d\theta}{\int_{-\infty}^{\infty} \phi(\theta - x) \left(\tau^2 + (\theta-\mu)^2\right)^{-1} d\theta} ;$$

where $\phi(\cdot)$ denotes the density of $N(0,1)$.

So, $E(\theta|x)$ is the ratio of expectations. $h_1(\theta) = \dfrac{\theta}{\tau^2 + (\theta-\mu)^2}$ to that of $h_2(\theta) = \dfrac{1}{\tau^2 + (\theta-\mu)^2}$ w.r.t. $N(x,1)$ distn.

So, we may simple sample $\theta_1, \theta_2, \ldots$ from $N(x,1)$ and we have

$$\hat{E}(\theta|x) = \frac{\sum_{i=1}^{m} \theta_i \left(\tau^2 + (\theta_i - \mu)^2\right)^{-1}}{\sum_{i=1}^{m} \left(\tau^2 + (\theta_i - \mu)^2\right)^{-1}}$$

as our Monte Carlo estimate of $E(\theta|x)$.

Another approach could be to write $E(\theta|x)$ as

$$E(\theta|x) = \int_{-\infty}^{\infty} \theta \exp\left(-(\theta - x)^2\right) \pi(\theta) d\theta \Big/ \int_{-\infty}^{\infty} \exp\left(-(\theta-x)^2\right) \pi(\theta) d\theta$$

when $\pi(\theta) = \text{Cauchy}(\mu, \tau)$.

A second Monte Carlo estimate is given by

$$\hat{E}(\theta|x) = \sum_{i=1}^{m} \theta_i \exp\left(-(\theta_i - x)^2\right) \Big/ \sum_{i=1}^{m} \exp\left(-(\theta_i - x)^2\right).$$

But one has to note that in either approach one doesn't necessarily get a sample, that would closely resemble a bonafide sample from the actual posterior distn.s.

Thus it is of need to find techniques of approximately generating samples from a distn. which is either the posterior distn. itself or a distn. resembling the posterior density closely.

Markov Chain Monte Carlo (MCMC):-

MCMC refers to a broad class of iterative procedure that produce random sequence with the Markov properties such that the Markov chain has the posterior distn. as its stationary distn. and the chain converges to that distn.. So, the technique is to produce a sequence X_t of RVs with such property. Lets first do a quick review of MC.

A prob. distn. π is called stationary or invariant for a transition prob. P or the associated MC $\{X_n\}$ if it is the case that when the prob. distn. of X_0 is π then the same is true for X_n for all $n \geq 1$.

Thus, in the countable state space case a prob distn. $\pi = \{\pi_i : i \in S\}$ is stationary for P, if $\forall\, j \in S$.

$$P(X_1 = j) = \sum_i P(X_1 = j \mid X_0 = i)\, P(X_0 = i)$$

$$= \sum_i \pi_i p_{ij} = P(X_0 = j) = \pi_j .$$

In vector notation, $\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots)$ satisfies $\underset{\sim}{\pi} = \underset{\sim}{\pi} P$.

A markov chain is called irreducible if for any two states i and j, $P(X_n = j \mid X_0 = i) > 0$ for some $n \geq 1$.

Theorem:- Let $\{X_n\}_{n \geq 0}$ be a MC with a countable state space S and a transition probability matrix P. Further suppose, it is irreducible with stationary prob. distn. $\pi = \{\pi_i : i \in S\}$. Then for any odd functn. $h : S \to \mathbb{R}$ and for any initial distn. of X_0. $\frac{1}{n} \sum_{i=0}^{n-1} h(X_i) \to \sum_j h(j)\, \pi_j$ in probability as $n \to \infty$.

To see how this is useful to us, note that given a prob. distn. π on S and a function h on S, suppose it is desired to find the "integral of h w.r.t. π" which reduces to $\sum_j h(j) \pi_j$ in the countable case.

Then look for an irreducible MC $\{X_n\}$ with space S and stationary distn π. Then starting from some initial value X_0, run the chain $\{X_j\}$ for a period of time, say, $0, 1, \ldots, n-1$ and consider as an estimate

$$\mu_n = \frac{1}{n} \sum_{i=0}^{n-1} h(X_j) \quad \text{and by the theorem}$$

$$\mu_n \approx \sum_j h(j)\pi_j \quad \text{for large } n.$$

This technique is called MCMC method.

An irreducible MC $\{X_n\}$ with a countable state space S is called aperiodic if for some $i \in S$, the gcd $\{n : p_{ii}^{(n)} > 0\} = 1$. Then, in addition to the theorem above, we have $\sum_j |P(X_n = j) - \pi_j| \to 0$ as $n \to \infty$.

for any initial distn. X_0. In other words, for large n, the distn. of X_n will be close to π_0.

Metropolis Hastings Algorithm:-

This is a very general MCMC method with wide applicability. The idea is not to directly simulate from given target density (which may be computationally very difficult) at all, but to simulate an easy Markov chain that has this target density as the density of its stationary distribution.

Let S be a finite (or countable) set. Let π be a prob. distn. of S. We shall call π, the target distribution. Let $Q = ((q_{ij}))$ be a transition prob. matrix such that for each i, it is computationally easy to generate a sample from the distn. $\{q_{ij} : j \in S\}$.

Let us generate MC. as follows:

If $X_n = i$ first sample from the distribution $\{q_{ij} : j \in S\}$ and denote that observation as Y_n. Thus choose X_{n+1} from the two values X_n and Y_n accordingly to

$$P(X_{n+1} = Y_n \mid X_n, Y_n) = \rho(X_n, Y_n)$$

$$P(X_{n+1} = X_n \mid X_n, Y_n) = 1 - \rho(X_n, Y_n).$$

where the "acceptance probability" $\rho(\cdot,\cdot)$ is given by

$$\rho(i,j) = \min\left\{\frac{\pi_j}{\pi_i} \cdot \frac{q_{ji}}{q_{ij}}, 1\right\} \text{ for all } (i,j) \text{ s.t. } \pi_i q_{ij} > 0.$$

Continuous Case:- Suppose S is a continuum and the target distribution $\pi(\,)$ has the density $p(\cdot)$. Then MH proceeds as follows:

Let Q be a transition function s.t. for each x, $Q(x,\cdot)$ has density $q(x,y)$. Then proceed as in the discrete case. But set the "acceptance probability" $\rho(x,y)$ to be

$$\rho(x,y) = \min\left\{\frac{p(y)}{p(x)} \frac{q(y,x)}{q(x,y)}, 1\right\}$$

for all (x,y) s.t. $p(x)q(x,y) > 0$.

Some popular choices of the proposal distn $q(\cdot,\cdot)$ are as follows:

In Metropolis et al. (1953), $q(\cdot,\cdot)$ was such that $q(x,y) = q(y,x)$ $\forall\ (x,y)$.

For example, when $\mathbf{X}$ is continuous $q(x,\cdot)$ might be a multivariate normal with mean X and constant covariance matrix Σ.

For the symmetric proposal distn the acceptance probability

$$\rho(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

A special case of M-H algorithm is the random walk Metropolis for which $q(x,y) = q(|x-y|)$, while choosing a proposal distn, its scale need to be chosen carefully.

A continuous proposal distn. generating small steps $Y - X_t$ will generally have a high acceptance rate but will nevertheless "mix" slowly i.e., move slowly around the support of the target $\pi(\cdot)$.

A bold proposal distribution generating large steps with the purpose moves from the body to the tails of the distn. giving small values of $\frac{\pi(y)}{\pi(x_t)}$ and a low probability of acceptance. Such a chain will frequently not move, again resulting in slow mixing.

The independence sampler (Tsirney, 1994) is a M-H algorithm where proposed $q(x,y) = q(y)$ doesn't depend on x. For this the acceptance prob. can be written in the form

$$\alpha(x,y) = \min\left(1, \frac{w(y)}{w(x)}\right), \text{ where } w(x) = \frac{\pi(x)}{q(x)}.$$

For the independence sampler to work well, $q(\cdot)$ should be a good approximation of $\pi(\cdot)$ but it is safest if $q(\cdot)$ is heavier tailed than $\pi(\cdot)$. To see this, suppose $q(\cdot)$ is highest tailed than $\pi(\cdot)$ and X_t is currently in the tails of $\pi(\cdot)$. Most candidates will not be in the tails so $w(X_t)$ will be much larger than $w(y)$ giving a low acceptance probability. The heavy tailed independent proposals help to avoid long periods stack in the tails.

Gibbs Sampling:- The Gibbs sampler is a technique suitable for generating a Markov chain of nice properties that has its stationary distribution a target distribution in higher dimensional space. The most interesting aspect of this technique is that to run this MC, it suffices to generate observations from univariate distns. Suppose you want to simulate a MC with $\pi(\underset{\sim}{\theta}|\underset{\sim}{y})$ where $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_K)$ as its stationary distribution. In Gibbs sampling in order to do this you proceed as follows:

Start with $\underset{\sim}{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0K})$

Generate θ_{11} from $\pi_1(\theta_1 | \theta_{02}, \ldots \theta_{0K}, \underset{\sim}{y})$

θ_{12} from $\pi_2(\theta_2 | \theta_{11}, \theta_{03}, \ldots, \theta_K, \underset{\sim}{y})$

$\vdots$

θ_{1K} from $\pi_K(\theta_K | \theta_{11}, \theta_{12}, \ldots, \theta_{1K-1}, \underset{\sim}{y})$

This gives $\underset{\sim}{\theta}_1 = (\theta_{11}, \theta_{12}, \ldots, \theta_{1K})$

Then produce θ_2 from $\pi_1(\theta_1 | \theta_{12}, \ldots, \theta_{1K}, \underset{\sim}{y})$ and so on.

Example:- (i) Suppose $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

$$X | Y = y \sim N(\rho y, 1-\rho^2)$$

$$Y | X = x \sim N(\rho x, 1-\rho^2)$$

To apply GS technique

1. Given X_i for X, draw $Y_i \sim N(\rho x_i, 1-\rho^2)$
2. Given Y_i for Y, draw $X_{i+1} \sim N(\rho y_i, 1-\rho^2)$.

(ii) $X | \theta \sim N(\theta, \sigma^2)$; σ^2 known.

$\theta \sim$ Cauchy (μ, τ)

$$\pi(\theta) \propto (\tau^2 + (\theta-\mu)^2)^{-1}$$

$$\propto \int_0^\infty \left(\frac{\lambda}{2\pi\tau^2}\right)^{1/2} \exp\left(-\frac{\lambda}{2\tau^2}(\theta-\mu)^2\right) \lambda^{1/2-1} \exp\left(-\frac{\lambda}{2}\right) d\lambda .$$

Thus $\pi(\theta)$ may be considered the marginal prior density from the joint prior density of (θ, λ), where, $\theta|\lambda \sim N(\mu, \tau^2/\lambda)$, and $\lambda \sim \text{Gamma}(1/2)$. It is clear that, this representation of $\pi(\cdot)$ leads to an implicit hierarchical prior structure with λ as a hyper parameter.

Consequently, $\pi(\theta|x)$ may be treated as the marginal density for $\pi(\theta, \lambda|x)$. To apply the GS note that

$$\theta|\lambda, x \sim N\left(\frac{\tau^2}{\tau^2+\lambda\sigma^2}\bar{x} + \frac{\lambda\sigma^2}{\tau^2+\lambda\sigma^2}\mu, \frac{\tau^2\sigma^2}{\tau^2+\lambda\sigma^2}\right)$$

$$\lambda|\theta, x \sim \lambda|\theta \sim \exp\left(\frac{\tau^2+(\theta-\mu)^2}{2\tau^2}\right).$$

Example: Suppose we are studying the distribution of the number of defectives X in the daily productions of a product. Consider the model $(x|y,\theta) \sim \text{Bin}(y,\theta)$, where, Y, a day's production, is a RV with Poisson distribution with unknown mean λ and θ is the probability (unknown) that a product is defective. Y is not observable and inference has to be made on the basis of X only. The prior distn. is such that $\theta|Y=y \sim \text{Beta}(\alpha, \gamma)$ with known α and γ independent of Y.

First note that $X|\theta \sim \text{Poisson}(\lambda\theta)$ (check)

since $\theta \sim \text{Beta}(\alpha, \gamma)$.

We have $\pi(\theta|X=x) \propto \exp(-\lambda\theta)\,\theta^{x+\alpha-1}(1-\theta)^{\gamma-1}$; $0<\theta<1$.

Note: $\theta|x$ is not a standard distn and hence posterior quantile can't be obtained in closed form. You can however use MCMC techniques. Let's see how Gibbs sampling works here. Instead of focusing $\theta|x$ directly view it as a marginal component of $(Y, \theta|x)$. It is easy to see that the full conditionals of this are given by

$$Y|x=x, \theta \sim x + \text{Poisson}(x(1-\theta))$$

and $$\theta|x=x, Y=y \sim \text{Beta}(\alpha+x, \gamma+y-x).$$

HIERARCHICAL BAYES (HB) APPROACH AND PARAMETRIC EMPIRICAL BAYES (PEB) APPROACH

We consider p similar but not identical populations with densities $f(x|\mu_1), \ldots, f(x|\mu_p)$. We illustrate below with normal densities:

Consider p independent random samples, each of size n, from p normal populations:

x_{ij}, $i = 1, 2, \ldots, n$ are iid $N(\mu_j, \sigma^2)$, $j = 1, 2, \ldots, p$.

For simplicity, assume σ^2 to be known. These p populations may correspond to p adjacent small areas with unknown per capita income $\mu_1, \ldots, \mu_p$, as in small area estimation. They could correspond to p clinical trials in a particular hospital with $\mu_1, \ldots, \mu_p$ being the mean effects of the drug being tested. It may be said that in these examples the different studied population are related to each other.

In order to assign a prior distribution for the μ_j's, we model them as exchangeable rather than i.i.d or just independent. An exchangeable, dependent structure is consistent with the assumption that the studies are similar in a broad sense, so they share many common elements. We use a standard way of generating exchangeable μ_j's which is as follows:

We introduce a hyperparameter $\eta = (\eta_1, \eta_2)$ and assume $\mu_1, \mu_2, \ldots, \mu_p$ are iid $N(\eta_1, \eta_2)$ given η and

$\eta = (\eta_1, \eta_2) \sim \pi(\cdot, \cdot)$ (some prior density for η)

$\underset{\sim}{X} = (x_{ij}, i = 1, \ldots, n, j = 1, \ldots, p)$

$\underset{\sim}{\mu} = (\mu_1, \ldots, \mu_p)$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$

Sufficient statistic $\underset{\sim}{\bar{X}} = (\bar{x}_1, \ldots, \bar{x}_p)$

Given η

$$(\underset{\sim}{X}, \underset{\sim}{\mu}) \sim f(\underset{\sim}{x}, \underset{\sim}{\mu} | \eta)$$

$$= \prod_{j=1}^{p} \left\{ \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_{ij} - \mu_j)^2 \right] \frac{1}{\sqrt{2\pi}\eta_2} \exp\left[-\frac{1}{2\eta_2} (\mu_j - \eta_1)^2 \right] \right\}$$

$$\pi(\underset{\sim}{\mu} | \underset{\sim}{x}, \underset{\sim}{\eta}) \propto \prod_{j=1}^{p} \left\{ \exp\left[-\frac{n}{2\sigma^2} (\mu_j - \bar{x}_j)^2 \right] \cdot \exp\left[-\frac{1}{2\eta_2} (\mu_j - \eta_1)^2 \right] \right\}$$

[Note that $\sum_{i=1}^{n} (x_{ij} - \mu_j)^2 = \sum_{i} (x_{ij} - \bar{x}_j)^2 + n(\bar{x}_j - \mu_j)^2$]

Thus, $(\mu_j | \underset{\sim}{x}, \underset{\sim}{\eta}) \equiv (\mu_j | \bar{x}_j, \eta)$ are independent normal with

$$E(\mu_j | \bar{x}_j, \eta) = \frac{\frac{n}{\sigma^2} \bar{x}_j + \eta_1/\eta_2}{\frac{n}{\sigma^2} + \frac{1}{\eta_2}} = \frac{\eta_2 \bar{x}_j + (\sigma^2/n)\eta_1}{\eta_2 + (\sigma^2/n)}$$

$$= (1 - B)\bar{x}_j + B\ \eta_1, \text{ where } B = \frac{\sigma^2/n}{\eta_2 + \frac{\sigma^2}{n}}.$$

$$\text{Var}(\mu_j | \bar{x}_j, \eta) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\eta_2}} = \frac{\eta_2 \sigma^2/n}{\eta_2 + \frac{\sigma^2}{n}}.$$

Given η, Bayes estimate of μ_j (w.r.t. square error loss) is

$$E(\mu_j | \underset{\sim}{x}, \eta) = E(\mu_j | \bar{x}_j, \eta) = (1 - B)\bar{x}_j + B\eta_1, \text{ where,}$$

$$B = \frac{\sigma^2/n}{\eta_2 + \sigma^2/n}.$$

but η is unknown.

Hierarchical Bayes Approach (HB):-

Consider a prior for $\eta = (\eta_1, \eta_2) \sim \pi(\cdot, \cdot)$ (prior density of η)

Find the posterior density $\pi(\eta | \underset{\sim}{x})$ of $\underset{\sim}{\eta}$.

$(\underset{\sim}{X}, \underset{\sim}{\mu}) \big|_{\underset{\sim}{\eta}} \sim f(\underset{\sim}{x}, \underset{\sim}{\mu} | \eta)$ as given above. Integrating out the μ_j's and holding η fixed, we get $\bar{x}_j$'s are independent and

$$\bar{x}_j | \eta \sim N\left(\eta_1, \eta_2 + \frac{\sigma^2}{n}\right)$$

This, together with the prior for η will give the posterior density $\pi(\eta | \underset{\sim}{x}) = \pi(\eta | \underset{\sim}{\bar{x}})$.

Then the Bayes estimate of μ_j is

$$E(\mu_j | \underset{\sim}{X}) = E(\mu_j | \bar{\underset{\sim}{X}}) = \int E(\mu_j | \bar{\underset{\sim}{X}}, \eta)\, \pi(\eta | \bar{\underset{\sim}{X}})\, d\eta$$

where, $E(\mu_j | \bar{\underset{\sim}{X}}, \eta) = E(\mu_j | \bar{X}_j, \eta) = (1-B)\bar{X}_j + B\eta_1$ and

$$B = \frac{\sigma^2/n}{\eta_2 + \sigma^2/n}.$$

Parametric Empirical Bayes Approach (PEB):-

Recall that given η, $\bar{X}_j$'s are independent and

$\bar{X}_j | \eta \sim f(\bar{x}_j | \eta) = N\left(\eta_1, \eta_2 + \frac{\sigma^2}{n}\right)$ density.

In this approach, one takes an intermediate position between a fully Bayes and a fully frequentist approach by treating the likelihood as given by $f(\bar{x}_j | \eta)$ obtained by integrating out the μ_j's. The μ_j's are treated as random variables as in Bayesian analysis whereas η_1, η_2 are treated as (fixed) unknown parameters as in frequentist analysis. The PEB approach differs from the fully Bayesian (HB) approach in that no prior is assigned to η, and η is estimated by MLE or by a suitable unbiased estimate.

Here, given η, $\bar{X}_1, \ldots, \bar{X}_p$ are iid $N\left(\eta_1, \eta_2 + \frac{\sigma^2}{n}\right)$

The best unbiased estimate (UMVUEs) of η_1 and B are given by $\hat{\eta}_1 = \frac{1}{p}\sum_{j=1}^{p} \bar{X}_j = \bar{\bar{X}}$, say and $\hat{B} = \frac{(p-3)\frac{\sigma^2}{n}}{S}$

$\left[\because S = \sum_{j=1}^{p} (\bar{X}_j - \bar{\bar{X}})^2\right]$

If $Y_1, Y_2, \ldots, Y_p$ are iid $N(\mu, \tau^2)$, UMVUE of μ is $\bar{Y} = \frac{1}{p}\sum Y_j$ and UMVUE of $\frac{1}{\tau^2}$ is $(p-3) \Big/ \sum_{j=1}^{p} (Y_j - \bar{Y})^2$.

Then PEB estimate of μ_j is $\hat{\mu}_j = (1-\hat{B})\bar{X}_j + \hat{B}\bar{\bar{X}}$

[a shrinkage estimate shrinking towards $\bar{\bar{X}}$]

If we assume $\eta_1 = 0$, i.e., we assume $\mu_1, \ldots, \mu_p$ are iid $N(0, \eta_2)$ then given η, $\bar{X}_1, \ldots, \bar{X}_p$ are iid $N\left(0, \eta_2 + \frac{\sigma^2}{n}\right)$.

Then UMVUE of B is $\hat{B} = \dfrac{(p-2)\frac{\sigma^2}{n}}{\sum_j \bar{X}_j^2}$ and the PEB estimate of μ_j is $\hat{\mu}_j = \left(1 - \dfrac{(p-2)\sigma^2/n}{\sum_j \bar{X}_j^2}\right)\bar{X}_j$, which is the well-known James - Stein estimate.

Remark:- The classical estimate of μ_j is $\bar{X}_j$, which is based only on the j^{th} sample. The Bayes estimate (HB or PEB) of μ_j, on the other hand, depends on $\bar{X}_j$ and also on $(\bar{X}_1, \ldots, \bar{X}_p)$, the full sufficient statistic because the posterior distribution of η or the estimate of η depends on all the $\bar{X}_j$'s. Thus the Bayes estimate learns from the full sufficient statistic justifying simultaneous estimation of all the μ_j's. This learning process is sometimes referred to as borrowing strength. If the modelling of μ_j's is realistic, we would expect the Bayes estimates to perform better than $\bar{X}_j$'s. that is what strikingly new in the case of large p, and follows from the exchangeability of the μ_j's.

Assignment I

1. Consider the problem of estimation of a real parameter θ with the loss function

$$L(\theta, a) = k_0(\theta - a) \quad \text{if } \theta - a \geq 0$$
$$= k_1(a - \theta) \quad \text{if } \theta - a < 0$$

S.T. the Bayes estimate is given by the quantile of order $\frac{k_0}{k_0+k_1}$ of the posterior distn.

Sol. $L(\theta, a) = k_0(\theta - a) - (k_0+k_1)(\theta - a) I(\theta < a)$ — ①

$= k_0(\theta - a) - (k_0+k_1)(\theta - a) I(\theta \leq a)$ — ②

Let $a_0 \in \mathbb{R}$ s.t. $P(\theta \leq a_0 \mid \underset{\sim}{X}) \geq \frac{k_0}{k_0+k_1}$

$P(\theta \geq a_0 \mid \underset{\sim}{X}) \geq \frac{k_1}{k_0+k_1}$.

We are to show $E_\theta\left[L(\theta, a) \mid \underset{\sim}{X}\right]$ is min. when $a = a_0$.

Assume $a < a_0$,

$L(\theta, a) - L(\theta, a_0)$

$= k_0(a_0 - a) - (k_0+k_1)(\theta - a) I(\theta < a) + (k_0+k_1)(\theta - a_0) I(\theta < a_0)$ [using ①]

$= k_0(a_0 - a) - (k_0+k_1)(\theta - a) I(\theta < a_0) + (k_0+k_1)(\theta - a) I(a \leq \theta < a_0)$
$+ (k_0+k_1)(\theta - a_0) I(\theta < a_0)$.

$= k_0(a_0 - a) + (k_0+k_1)(a - a_0) I(\theta < a_0) + \underbrace{(k_0+k_1)(\theta - a) I(a < \theta < a_0)}_{\geq 0}$

$\geq (a_0 - a)(k_0+k_1)\left[\frac{k_0}{k_0+k_1} - I(\theta < a_0)\right]$

$E\left[L(\theta, a) - L(\theta, a_0) \mid \underset{\sim}{X}\right]$

$\geq (a_0 - a)(k_0+k_1)\left[\frac{k_0}{k_0+k_1} - P(\theta < a_0 \mid \underset{\sim}{X})\right] > 0$

Assume $a > a_0$

$$L(\theta,a) - L(\theta,a_0)$$

$$= K_0(a_0-a) - (K_0+K_1)(\theta-a) I(\theta \le a) + (K_0+K_1)(\theta-a_0) I(\theta \le a_0) \quad [\text{Using } (2)]$$

$$= K_0(a_0-a) - (K_0+K_1)(\theta-a) I(\theta \le a_0) - (K_0+K_1)(\theta-a) I(a_0 \le \theta \le a) + (K_0+K_1)(\theta-a_0) I(\theta \le a_0)$$

$$= (K_0+K_1)(a-a_0)\left[I(\theta \le a_0) - \frac{K_0}{K_0+K_1} \right] + \underbrace{(K_0+K_1)(\theta-a) I(a_0 < \theta \le a)}_{\ge 0}$$

$$\ge (K_0+K_1)(a-a_0)\left[I(\theta \le a_0) - \frac{K_0}{K_0+K_1} \right]$$

$$\therefore E\left[L(\theta,a) - L(\theta,a_0) \,\middle|\, \underset{\sim}{X} \right]$$

$$\ge (K_0+K_1)(a-a_0)\left[P(\theta \le a_0 \,|\, \underset{\sim}{X}) - \frac{K_0}{K_0+K_1} \right]$$

$$\ge 0.$$

$$\therefore a_0 = \arg\min_a E\left(L(\theta,a) \,|\, \underset{\sim}{X} \right).$$

4. S.T. the result on asymptotic normality of the posterior distn. of $\sqrt{n}(\theta - \hat{\theta}_n)$, proved in the class, implies consistency of the posterior distn. of θ at θ_0.

Sol. We know that with $[P_{\theta_0}]$ probability 1

$$\sqrt{n}\left(\theta - \hat{\theta}_n\right) \,|\, \underset{\sim}{X}_n \xrightarrow{d} N\left(0, I^{-1}(\theta_0)\right)$$

$\hat{\theta}_n$ is strongly consistent soln of $L_n'(\theta) = 0$.

$$\hat{\theta}_n \to \theta_0 \text{ with } [P_{\theta_0}] \text{ prob. } 1.$$

So, $\hat{\theta}_n \,|\, \underset{\sim}{X}_n \to \theta_0$ a.s. $[P_{\theta_0}]$

$$\sqrt{n}\left(\theta - \hat{\theta}_n\right) \,|\, \underset{\sim}{X}_n \xrightarrow{d} N\left(0, I^{-1}(\theta_0)\right) \text{ a.s. } [P_{\theta_0}]$$

$$\Rightarrow \left(\theta - \hat{\theta}_n\right) \,|\, \underset{\sim}{X}_n \xrightarrow{P} 0 \text{ a.s. } [P_{\theta_0}]$$

$$\left(\hat{\theta}_n - \theta_0\right) \,|\, \underset{\sim}{X}_n \xrightarrow{P} 0 \text{ a.s. } [P_{\theta_0}]$$

$$\Rightarrow (\theta - \theta_0) \,|\, \underset{\sim}{X}_n \xrightarrow{P} 0 \text{ a.s. } [P_{\theta_0}]$$

$$\therefore \theta \,|\, \underset{\sim}{X}_n \xrightarrow{P} \theta_0 \text{ a.s. } [P_{\theta_0}]$$

$\Rightarrow$ For any U open set containing θ_0 $P[\theta \in U \,|\, X_n] \to 1$ a.s. $[P_{\theta_0}]$

$\Rightarrow \theta \,|\, X_n$ is consistent at θ_0.

5. Show that condition (A4), used to prove asymptotic normality of posterior distn, holds when $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$.

Sol. A4 condition: for any $\delta > 0$,

$$P_{\theta_0}\left[\text{For some } \epsilon > 0, \sup_{|\theta-\theta_0|>\delta} \frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right] < -\epsilon \ \forall n\right] = 1.$$

For $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$

$$\frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right] = -\frac{1}{n}\left[\sum_{i=1}^{n}\left[(X_i-\theta)^2 - (X_i-\theta_0)^2\right]\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}(2X_i - \theta - \theta_0)(\theta_0 - \theta)$$

$$= -(\theta_0-\theta)(2\bar{X}_n - \theta - \theta_0)$$

$$\xrightarrow{a.s.} -(\theta_0-\theta)(2\theta_0 - \theta - \theta_0)$$

$$= -(\theta_0-\theta)^2 \text{ under } [P_{\theta_0}].$$

For $\delta > 0$, for $\omega \in \Omega$ s.t.

$$\frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right](\omega) \longrightarrow -(\theta-\theta_0)^2$$

then $\sup_{|\theta-\theta_0|>\delta} \frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right](\omega) \xrightarrow{a.s.} \sup_{|\theta-\theta_0|>\delta} -(\theta-\theta_0)^2$

$$= -\delta^2.$$

choose $\epsilon = \frac{\delta^2}{2} > 0$ then $\forall n$

$$\frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right](\omega) < -\epsilon.$$

As, $P_{\theta_0}\left[\omega \in \Omega : \frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right](\omega) \longrightarrow -(\theta-\theta_0)^2\right] = 1$

So, $P_{\theta_0}\left[\exists\ \epsilon > 0, \sup_{|\theta-\theta_0|>\delta} \frac{1}{n}\left[L_n(\theta) - L_n(\theta_0)\right] < -\epsilon \ \forall n\right] = 1.$

10. (a) Let $X_1, \ldots, X_n$ be iid $\sim N(\theta, 1)$. Consider the problem of testing $H_0: \theta \le \theta_0$ Vs. $H_1: \theta > \theta_0$. A classical test rejects H_0 if $T = \sqrt{n}(\bar{X} - \theta_0)$ is large. Let t be the observed value of T. Find the p-value (in terms of t). Consider now the uniform prior $\pi(\theta) \equiv 1$. Find the posterior prob. of H_0 (in terms of t) and compare it with the p-value.

Soln:- $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$

$\bar{X}_n \sim N(\theta, 1/n)$ under $[P_\theta]$

$\sqrt{n}(\bar{X}_n - \theta) \sim N(0,1)$ under $[P_\theta]$

$$P_\theta\left[\sqrt{n}(\bar{X} - \theta_0) > t\right] = P_\theta\left[\sqrt{n}(\bar{X} - \theta) - \sqrt{n}(\theta_0 - \theta) > t\right]$$

$$= P_\theta\left[\sqrt{n}(\bar{X} - \theta) > t + \sqrt{n}(\theta_0 - \theta)\right]$$

$$= 1 - \Phi\left(t + \sqrt{n}(\theta_0 - \theta)\right).$$

$$p = \sup_{\theta \le \theta_0}\left(1 - \Phi\left(t + \sqrt{n}(\theta_0 - \theta)\right)\right)$$

$$= 1 - \sup_{\theta \le \theta_0} \Phi\left(t + \sqrt{n}(\theta_0 - \theta)\right)$$

$$= 1 - \Phi(t).$$

2nd Part

Alt:

$$P_{\theta|\underline{x}}(\theta \le \theta_0)$$

$$= P\left(N\left(\bar{x}, \frac{\sigma^2}{n}\right) \le \theta_0\right)$$

$$= P\left(N(0,1) \le \frac{\sqrt{n}(\theta_0 - \bar{x})}{\sigma}\right)$$

$$= P(N(0,1) \le -t)$$

$$= \Phi(-t)$$

Consider $\pi(\theta) \equiv 1$.

$$f(x_1, \ldots, x_n | \theta) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

$$\propto \exp\left[-\frac{n\theta^2}{2} + n\theta\bar{X}_n\right]$$

$$\propto \exp\left[-\frac{n}{2}(\theta - \bar{X}_n)^2\right]$$

$$\pi(\theta | x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n | \theta)\pi(\theta)$$

$$\propto \exp\left[-\frac{n}{2}(\theta - \bar{X}_n)^2\right]$$

So, $\theta | x_1, \ldots, x_n \sim N\left(\bar{X}_n, \frac{1}{n}\right)$.

$$P(H_0 | x_1, \ldots, x_n) = P(\theta \le \theta_0 | x_1, \ldots, x_n)$$

$$= P\left(\sqrt{n}(\theta - \bar{X}_n) \le \sqrt{n}(\theta_0 - \bar{X}_n) \,\middle|\, x_1, \ldots x_n\right)$$

$$= \Phi\left(\sqrt{n}(\theta - \bar{X}_n)\right) = \Phi(-t)$$

$$= 1 - \Phi(t).$$

(11) Let the sample space be $\{1,2,\ldots,k\}$ and $P=(p_1,\ldots,p_k)$ be a random prob. distn. on this sample space. Let $X_1,X_2,\ldots,X_n$ be iid $\sim P$, and $P\sim$ Dirichlet $(\alpha_1,\ldots,\alpha_k)$, $\alpha_i>0\ \forall i$. S.T. any subset A of the sample space, the posterior mean of $P(A)$ is a weighted average of its prior mean and $P_n(A)$ where, P_n denotes the empirical distn of $X_1,X_2,\ldots,X_n$.

Sol. If $P\sim D(\alpha_1,\ldots,\alpha_k)$; $\alpha_i>0\ \forall i$
and $A\subseteq \mathcal{X}=\{1,2,\ldots,k\}$

then $(p_1,\ldots,p_k)\sim D(\alpha_1,\ldots,\alpha_k)$

$P(A)=\sum_{i=1}^{k} p_i I(i\in A)\sim \text{Beta}(\alpha(A),\ \alpha(\mathcal{X})-\alpha(A))$.

where $\alpha(A)=\sum_{i=1}^{k}\alpha_i I(i\in A)$ [follows from the properties of Dirichlet distn.]

So, $E(P(A))=\frac{\alpha(A)}{\alpha(\mathcal{X})}=\bar{\alpha}(A)$.

Let $X_1,\ldots,X_n\sim P$

then $f(X_1,\ldots,X_n\mid P)=p_1^{n_1}\cdots p_k^{n_k}$ where $n_j=\sum_{i=1}^{n} I(X_i=j)$

$P\sim D(\alpha_1,\ldots,\alpha_k)$

$\Rightarrow \pi(P)\propto p_1^{\alpha_1-1}\cdots p_k^{\alpha_k-1}$

$\pi(P\mid X_1,\ldots,X_n)\propto \pi(P)\,f(X_1,\ldots,X_n\mid P)$

$=p_1^{n_1+\alpha_1-1}\cdots p_k^{n_k+\alpha_k-1}$

So, $P\mid X_1,\ldots,X_n\sim D(\alpha_1+n_1,\ldots,\alpha_k+n_k)$

Define $\lambda:\mathcal{P}(\mathcal{X})\to\mathbb{R}^+$

Set $\lambda(\{i\})=n_i$ for $i\in\mathcal{X}$.

Then $E(P(A)\mid X_1,\ldots,X_n)=\frac{\alpha(A)+\lambda(A)}{\alpha(\mathcal{X})+n}$.

$=\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}\cdot\frac{\alpha(A)}{\alpha(\mathcal{X})}+\frac{n}{\alpha(\mathcal{X})+n}\cdot\frac{\lambda(A)}{n}$

$\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}\cdot\frac{\alpha(A)}{\alpha(\mathcal{X})}$ $\uparrow$ $E(P(A))$; $\frac{n}{\alpha(\mathcal{X})+n}\cdot\frac{\lambda(A)}{n}$ $\uparrow$ $P_n(A)$

(8) Let X_1, X_2 be a r.s. from a $U(\theta - 1/2, \theta + 1/2)$ distn, $-\infty < \theta < \infty$. Consider the noninformative prior $\pi(\theta) = 1$. What will be your 95% credible interval for θ?

Sol. $\pi(\theta) = 1$.

$$f(x_1, x_2 | \theta) = I\left(\theta - 1/2 < x_{(1)} \le x_{(2)} < \theta + 1/2\right)$$

$$\pi(\theta | x_1, x_2) \propto I\left(\theta - 1/2 < x_{(1)} \le x_{(2)} < \theta + 1/2\right)$$

$$\Rightarrow \quad \theta < x_{(1)} + 1/2 \quad \text{and} \quad \theta > x_{(2)} - 1/2.$$

Also, $x_{(2)} - 1/2 < \theta < x_{(1)} + \frac{1}{2}$.

$$\therefore \pi(\theta | x_1, x_2) = I\left(x_{(2)} - 1/2 < \theta < x_{(1)} + 1/2\right)$$

~~$= 1 - |x_{(1)} - x_{(2)}|$~~ $= 1 - |x_1 - x_2|$

95% credible set is any

$$C \subseteq \left(x_{(2)} - 1/2, x_{(1)} + 1/2\right)$$

s.t. $P(\theta \in C | x_1, x_2) = 0.95$

Let $C = (\bar{x} - k, \bar{x} + k)$ then

$$P(\theta \in C | x_1, x_2) = \frac{2k}{1 - |x_1 - x_2|} = 0.95$$

$$\Rightarrow k = \frac{0.95\left(1 - |x_1 - x_2|\right)}{2}.$$

So, $C = \left(\bar{x} \mp \frac{0.95\left(1 - |x_1 - x_2|\right)}{2}\right)$.