Prior Information and Subjective Probability

u892603　郭彥甫

Outline

- Subjective Probability
- Subjective Determination of the Prior Density
- Noninformative Priors ▶
- Maximum Entropy Priors ▶
- Using the Marginal Distribution to Determine the Prior ▶
- Hierarchical Prior ▶
- Criticisms

Subjective Probability

- Prior information
- Classical concept of probability:

  → frequency viewpoint
- Subjective probability:

  → deal with random θ that frequency viewpoint does not apply
- Ex: coin tossing & unemployment rate for next year

Subjective Determination of the Prior Density

- The Histogram Approach
- The Relative Likelihood Approach
- Matching a Given Functional Form
- CDF Determination

The Histogram Approach

- When Θ is an interval of real line, the most approach to use is the histogram. Divide Θ into intervals, determine the subjective probability of each interval, and plot a probability histogram.
- Short cut: how many time interval?

  what size of intervals?

The Relative Likelihood Approach

- When Θ is a subset of the real line, compare the intuitive likelihoods of various points in, and sketch a prior density.

Ex: $\Theta=[0,1]$

Determine the most likely parameter point $\theta = \frac{3}{4}$, which is three times as likely as $\theta = 0$, the least likely ones. Then determine three other points compared with $\theta = 0$ and sketch the result.

Matching a Given Functional Form

- Assume that $\pi(\theta)$ is of a given functional form, and choose the density which most closely matched prior beliefs
- After determined the functional form, choose parameters for the function

$\longrightarrow$ from estimated prior moments.

$\longrightarrow$ subjectively estimate several fractiles of prior distribution, and matching these fractiles

Draw backs: Only useful when certain specific functional forms of prior are assumed.

Matching a Given Functional Form

- Example:

$\Theta=(-\infty, \infty)$, prior is thought to be from normal family. Determine the median is 0, and the quartiles are -1 and 1. Since mean is equal to median, $\mu=0$.

$\because P(Z<-1/(2.19)^{1/2})=1/4$ when Z is N(0,1).

$\therefore$ the density of prior is N(0,2.19)

CDF Determination

- This approach can be done by subjectively determining several α-fractiles, $z(\alpha)$, plotting the points $(z(\alpha), \alpha)$, and sketching a smooth curve joining them.

Discussion

- Multivariate prior density can be considerable

$\longrightarrow$ The easiest way is the use of a given functional form, then only a few parameters need to be determined subjectively. Also, more easier is the case in which the coordinate, θ_i, of θ are thought to be independent. The prior is then the product of the univariate prior density of the θ_i.

Ex: $\prod \pi(\theta_i)= \pi(\theta_1, \theta_2)$

If not, the best way is to determine conditional and marginal prior densities

Ex: $\pi(\theta_1,\theta_2)=\pi(\theta_1)\ \pi(\theta_2 \mid \theta_1)$

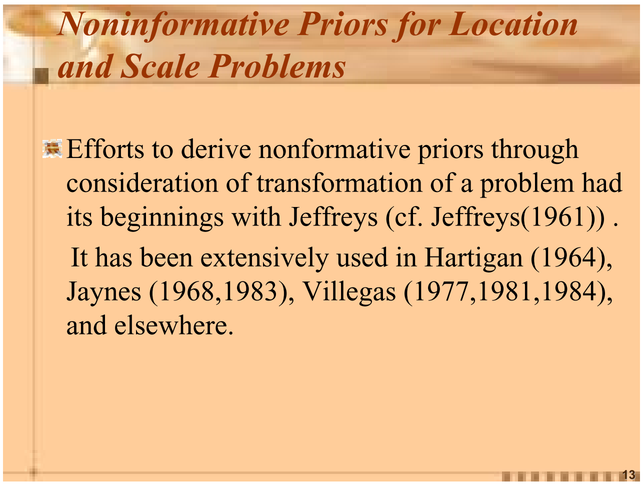
Noninformative Priors

- Because of the compelling reasons to perform a conditional analysis and the attractiveness of using Bayesian machinery to do so, there have been attempts to use the Bayesian approach even when no prior information is available.

Ex: Suppose the parameter of interest is a normal mean θ, so $\Theta=(-\infty, \infty)$. Noninformative prior is chosen to be $\pi(\theta)=1$ (not $\pi(\theta)=c>0$) (called the uniform density on R^1, and was introduce by Laplace(1892))

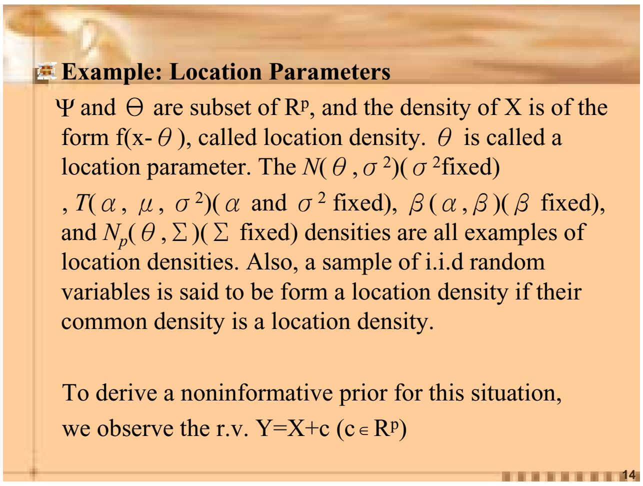
Noninformative Priors

- Sometimes, noninformative cannot maintain consistency.
- The lack of invariance of the constant prior has led to a search for noninformative priors which are appropriately invariant under transformations.



Noninformative Priors for Location and Scale Problems

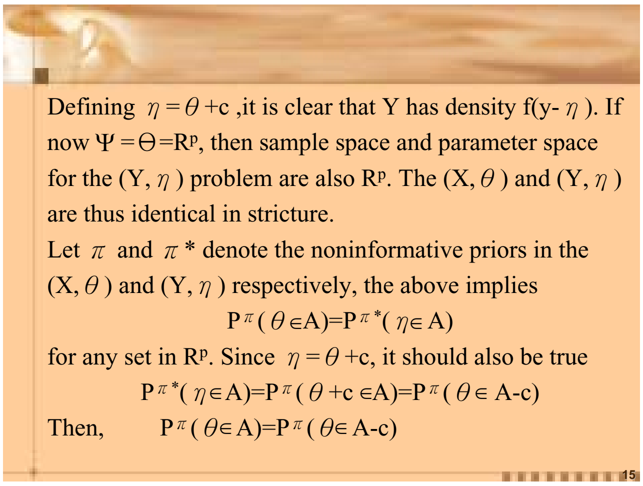
- Efforts to derive nonformative priors through consideration of transformation of a problem had its beginnings with Jeffreys (cf. Jeffreys(1961)) . It has been extensively used in Hartigan (1964), Jaynes (1968,1983), Villegas (1977,1981,1984), and elsewhere.



Example: Location Parameters

Ψ and Θ are subset of R^p, and the density of X is of the form f(x- θ), called location density. θ is called a location parameter. The $N(\theta, \sigma^2)(\sigma^2$fixed) , $T(\alpha, \mu, \sigma^2)(\alpha$ and σ^2 fixed), $\beta(\alpha, \beta)(\beta$ fixed), and $N_p(\theta, \Sigma)(\Sigma$ fixed) densities are all examples of location densities. Also, a sample of i.i.d random variables is said to be form a location density if their common density is a location density.

To derive a noninformative prior for this situation, we observe the r.v. Y=X+c ($c \in R^p$)



Defining $\eta = \theta$ +c ,it is clear that Y has density f(y- η). If now $\Psi = \Theta = R^p$, then sample space and parameter space for the (Y, η) problem are also R^p. The (X, θ) and (Y, η) are thus identical in stricture.

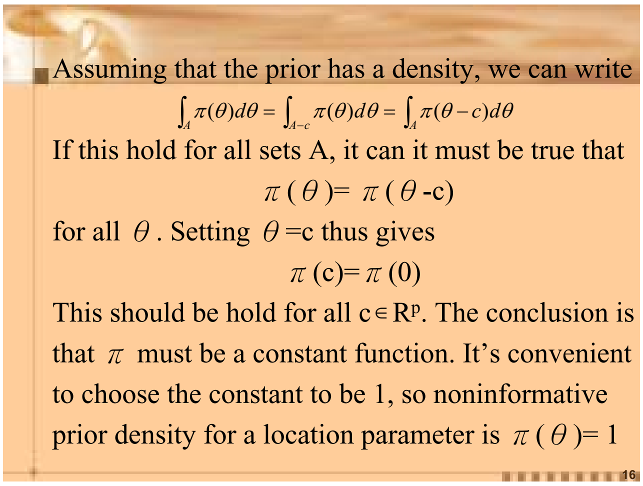
Let π and π * denote the noninformative priors in the (X, θ) and (Y, η) respectively, the above implies

$$P^{\pi}(\theta \in A)=P^{\pi*}(\eta \in A)$$

for any set in R^p. Since $\eta = \theta$ +c, it should also be true

$$P^{\pi*}(\eta \in A)=P^{\pi}(\theta + c \in A)=P^{\pi}(\theta \in A\text{-}c)$$

Then, $P^{\pi}(\theta \in A)=P^{\pi}(\theta \in A\text{-}c)$



Assuming that the prior has a density, we can write

$$\int_A \pi(\theta)d\theta = \int_{A-c} \pi(\theta)d\theta = \int_A \pi(\theta - c)d\theta$$

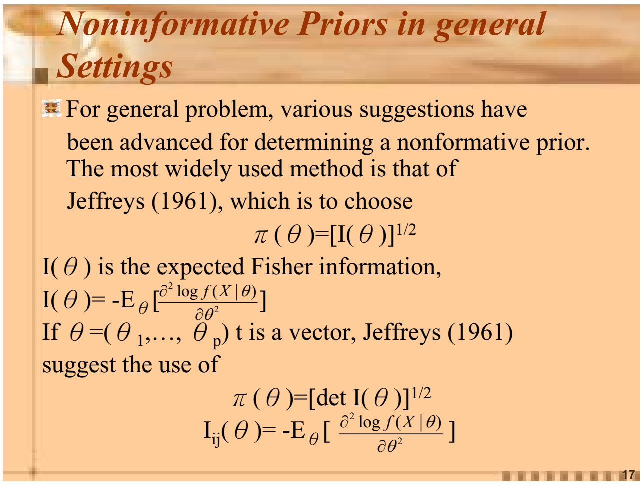
If this hold for all sets A, it can it must be true that

$$\pi(\theta) = \pi(\theta\text{-c})$$

for all θ . Setting θ =c thus gives

$$\pi(c) = \pi(0)$$

This should be hold for all $c \in R^p$. The conclusion is that π must be a constant function. It's convenient to choose the constant to be 1, so noninformative prior density for a location parameter is $\pi(\theta)$= 1



Noninformative Priors in general Settings

- For general problem, various suggestions have been advanced for determining a nonformative prior. The most widely used method is that of Jeffreys (1961), which is to choose

$$\pi(\theta)=[I(\theta)]^{1/2}$$

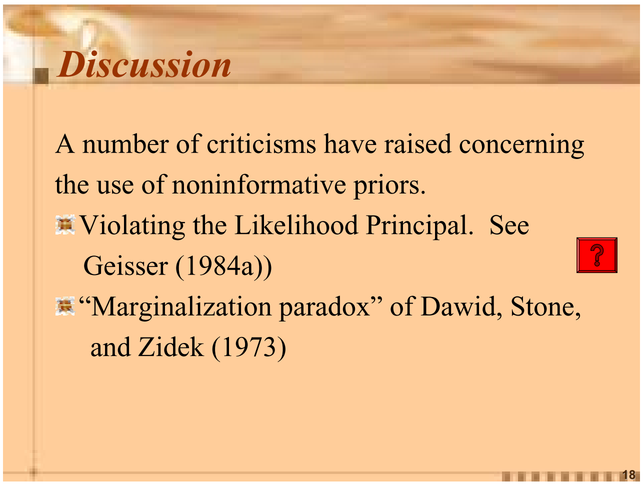
I(θ) is the expected Fisher information,

$I(\theta)= -E_{\theta}[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}]$

If $\theta = (\theta_1, \ldots, \theta_p)$ t is a vector, Jeffreys (1961) suggest the use of

$$\pi(\theta)=[\det I(\theta)]^{1/2}$$

$$I_{ij}(\theta)= -E_{\theta}[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}]$$



Discussion

A number of criticisms have raised concerning the use of noninformative priors.

- Violating the Likelihood Principal. See Geisser (1984a))
- "Marginalization paradox" of Dawid, Stone, and Zidek (1973)

Discussion

There are two common responses to these criticisms of noninformative prior Bayesian analysis.

- The first response, attempted by some noninformative prior Bayesians, is to argue for the "correctness" of their favorite noninformative prior approach, together with attempts to rebut the "paradoxes" and "counterexamples."
- The second response is to argue that, operationally, it is rare for the choice of a noninformative prior to markedly affect the answer, so that any reasonable noninformative prior can be used.

Maximum Entropy Priors

- Frequently partial prior information is available, outside of which it is desired to use a prior that is as noninformative as possible
- **Definition 1:** Assume Θ is discrete, let π be a probability density on Θ. The entropy of π, to be denoted

$$\xi(\pi) = -\sum_{\Theta} \pi(\theta_i) \log \pi(\theta_i)$$

⇨Entropy has a direct relationship to information theory, and in a sense measures the amount of uncertainty inherent in the probability distribution

Assume that partial prior information concerning θ is available.

$E^{\pi}[g_k(\theta)] = \sum_i \pi(\theta_i) g_k(\theta_i) = \mu_k$, k=1,…,m *

It seems reasonable to seek the prior distribution which maximizes entropy among all those distributions which satisfy the given set of restrictions. The solution is given by

$$\bar{\pi}(\theta_i) = \frac{\exp\left\{\Sigma_{k=1}^{m} \lambda_k g_k(\theta_i)\right\}}{\Sigma_i \exp\left\{\Sigma_{k=1}^{m} \lambda_k g_k(\theta_i)\right\}},$$

##proof

where λ_k are constants to be determined form the constraint in *

If Θ is continuous, the use of maximum entropy becomes more complicated. Jaynes (1968) makes a stronger case for defining entropy as

$\xi(\pi) = -E^{\pi}[\log \frac{\pi(\theta)}{\pi_0(\theta)}] = -\int \pi(\theta) \log(\frac{\pi(\theta)}{\pi_0(\theta)}) d\theta$,where $\pi_0(\theta)$ is the natural "invariant" noninformative prior for the problem.

In the presence of partial prior information of the form

$E^{\pi}[g_k(\theta)] = \int_{\Theta} g_k(\theta)\pi(\theta) d\theta = \mu_k$, k=1,…,m, **

the prior density which maximizes $\xi(\pi)$ is given by

$$\bar{\pi}(\theta) = \frac{\pi_0(\theta) \exp\left\{\Sigma_{k=1}^{m} \lambda_k g_k(\theta)\right\}}{\int_{\Theta} \pi_0(\theta)_i \exp\left\{\Sigma_{k=1}^{m} \lambda_k g_k(\theta)\right\}},$$

where λ_k are constants to be determined form the constraint in **

Example:

Assume $\Theta = R^1$, θ is a location parameter. The natural noninformative prior is then $\pi_0(\theta) = 1$. It is believed that The true prior mean is μ and variance is σ^2. These restriction are of the form ** with $g_1(\theta) = \theta$, $\mu_1 = \mu$, $\mu_2 = \sigma^2$,and $g_2(\theta) = (\theta - \mu)^2$

The maximum entropy prior, subject to these restriction is

$$\bar{\pi}(\theta) = \frac{\exp\left[\lambda_1\theta + \lambda_2(\theta-\mu)^2\right]}{\int_{\Theta} \exp\left[\lambda_1\theta + \lambda_2(\theta-\mu)^2\right] d\theta},$$

where λ_1 and λ_2 are to be chosen from **. Clearly

$\lambda_1\theta + \lambda_2(\theta-\mu)^2 = \lambda_2[\theta - \frac{\lambda_1}{2\lambda_2}]^2 + [\lambda_1\mu - \frac{\lambda_1^2}{4\lambda_2}]$

Example (conti)

Hence

$$\bar{\pi}(\theta) = \frac{\exp\left\{[\lambda_2[\theta - (\mu - \lambda_1/2\lambda_2)]^2\right\}}{\int_{-\infty}^{\infty} \exp\left\{[\lambda_2[\theta - (\mu - \lambda_1/2\lambda_2)]^2\right\} d\theta}$$

The denominator is a constant, so $\bar{\pi}(\theta)$ is normal density with mean $\mu - \lambda_1/2\lambda_2$ and variance $-1/2\lambda_2$. Choose $\lambda_1=0$ and $\lambda_2=-1/2\sigma^2$ satisfies **. Thus $\bar{\pi}(\theta)$ is a $N(\mu, \sigma^2)$ density.

Difficulties arising from this approach:

Although the need to use a noninformative prior in the derivation of $\bar{\pi}$ is not too serious, a more serious problem is that often $\bar{\pi}$ won't exist.

Using the Marginal Distribution to Determine the Prior

- If X has probability density f(x| θ), and θ has probability density $\pi(\theta)$, then the joint density of X and θ is h(x, θ)=f(x| θ) $\pi(\theta)$.
- **Definition 2:** The marginal density of X is

$$m(x|\pi)=\int_{\Theta} f(x|\theta)dF^{\pi}(\theta)=\begin{cases}\int_{\Theta} f(x|\theta)\pi(\theta)d\theta & \text{(conti case)}\\ \sum_{\Theta} f(x|\theta)\pi(\theta) & \text{(discrete case)}\end{cases}$$

- Bayesians have long used m to check assumptions. If m (for the actual observed data x) turns out to be small, then the assumptions (the model f and prior π) have not "predicted" what actually occurred and are suspect.

Information About m

- Subjective knowledge
- The data itself

The ML- ∏ Approach to Prior Selection

- In Definition 2, it was pointed out that m(x| π) reflects the plausibility of f and π , in the light of the data. If we treat f as definitely known, it follows that m(x| π) reflects the plausibility of π
- It is reasonable to consider m(x| π) as a likelihood function for π . Faced with a "likelihood function" for π , a natural method of choosing π is to use maximum likelihood.

The ML- ∏ Approach to Prior Selection (conti)

- **Definition 3:** Suppose Γ is a class of priors under consideration, and that $\pi^* \in \Gamma$ satisfies (for the observed data x)

$$m(x|\pi^*)=\sup_{\pi\in\Gamma} m(x|\pi)$$

Then π^* will be called type ∏ maximum likelihood prior, or ML- ∏ prior for short.

- When Γ is the class $\Gamma=\{\pi:\ \pi(\theta)=g(\theta|\ \lambda),\ \lambda\in\Lambda\}$,

then

$$\sup_{\pi\in\Gamma} m(x|\pi)=\sup_{\lambda\in\Lambda} m(x|\ g(\theta|\ \lambda)),$$

so that one simply max over the hyperparameter λ .

Hierarchical Prior

- Hierarchical Prior also called a multistage prior. The idea is that one may have structural and subjective prior information at the same time, and it is often convenient to model this in stages. For instance, in the Bayes scenario, structural knowledge that the θ_i were i.i.d. led to the first stage prior description

$$\pi_1(\theta)=\prod_{i=1}^{p}\pi_0(\theta_i)$$

The hierarchical approach would seek to place a "second stage" subjective prior on π_0.

- The hierarchical approach is most commonly used when the first stage, Γ, consists of priors of a certain functional form.

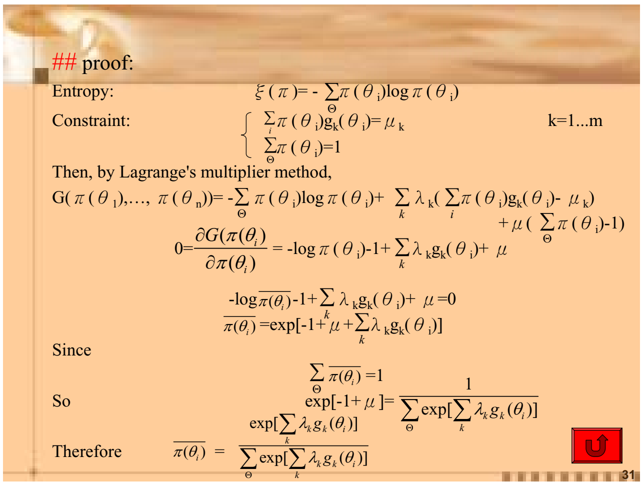
Criticisms

- **Objectivity**

⇨ Classical statistics is "objective" and hence suitable for the needs of science, while Bayesians is "subjective" and only useful for making personal decisions.

- **Misuse of prior distributions**
- **Robustness** (in section 4.7)
- **Data or model dependent priors**

⇨ The idealized Bayesian view is that θ is a quantity about which separate information exists, and that this information is to be combined with that in the data. The approach presumes the prior doesn't depend in any way on the data.



## proof:

Entropy: $\xi(\pi) = -\sum_{\Theta}\pi(\theta_i)\log\pi(\theta_i)$

Constraint: $\begin{cases}\sum_i \pi(\theta_i)g_k(\theta_i)=\mu_k & k=1\ldots m\\ \sum_{\Theta}\pi(\theta_i)=1\end{cases}$

Then, by Lagrange's multiplier method,

$$G(\pi(\theta_1),\ldots,\pi(\theta_n)) = -\sum_{\Theta}\pi(\theta_i)\log\pi(\theta_i)+\sum_k\lambda_k\Big(\sum_i\pi(\theta_i)g_k(\theta_i)-\mu_k\Big)+\mu\Big(\sum_{\Theta}\pi(\theta_i)-1\Big)$$

$$0=\frac{\partial G(\pi(\theta_i)}{\partial\pi(\theta_i)} = -\log\pi(\theta_i)-1+\sum_k\lambda_k g_k(\theta_i)+\mu$$

$$-\log\overline{\pi(\theta_i)}-1+\sum_k\lambda_k g_k(\theta_i)+\mu=0$$

$$\overline{\pi(\theta_i)}=\exp[-1+\mu+\sum_k\lambda_k g_k(\theta_i)]$$

Since

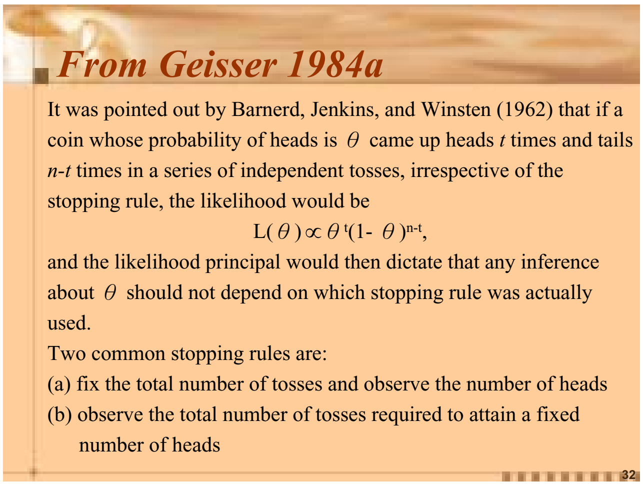
$$\sum_{\Theta}\overline{\pi(\theta_i)}=1$$

So

$$\exp[-1+\mu]=\frac{1}{\sum_{\Theta}\exp[\sum_k\lambda_k g_k(\theta_i)]}$$

Therefore

$$\overline{\pi(\theta_i)} = \frac{\exp[\sum_k\lambda_k g_k(\theta_i)]}{\sum_{\Theta}\exp[\sum_k\lambda_k g_k(\theta_i)]}$$



From Geisser 1984a

It was pointed out by Barnerd, Jenkins, and Winsten (1962) that if a coin whose probability of heads is θ came up heads *t* times and tails *n-t* times in a series of independent tosses, irrespective of the stopping rule, the likelihood would be

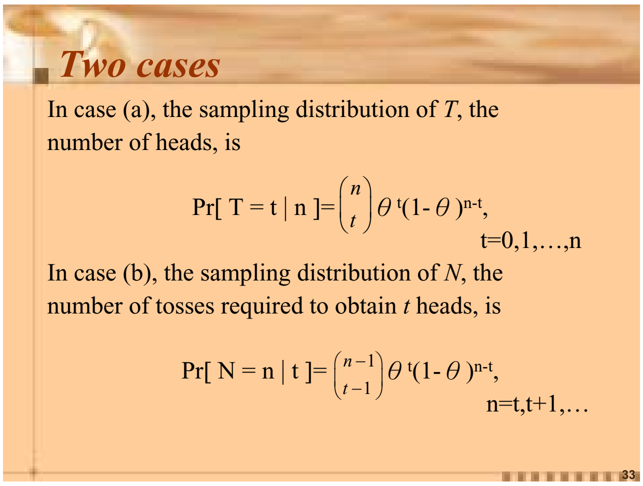
$$L(\theta)\propto\theta^{t}(1-\theta)^{n-t},$$

and the likelihood principal would then dictate that any inference about θ should not depend on which stopping rule was actually used.

Two common stopping rules are:

(a) fix the total number of tosses and observe the number of heads

(b) observe the total number of tosses required to attain a fixed number of heads



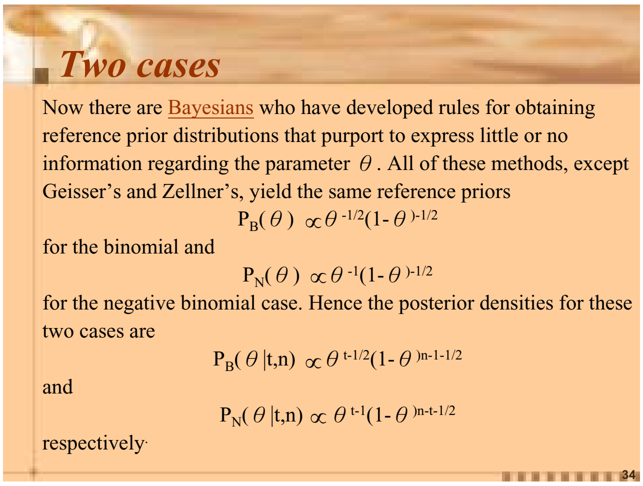
Two cases

In case (a), the sampling distribution of *T*, the number of heads, is

$$\Pr[\,T=t\mid n\,]=\binom{n}{t}\theta^{t}(1-\theta)^{n-t},\quad t=0,1,\ldots,n$$

In case (b), the sampling distribution of *N*, the number of tosses required to obtain *t* heads, is

$$\Pr[\,N=n\mid t\,]=\binom{n-1}{t-1}\theta^{t}(1-\theta)^{n-t},\quad n=t,t+1,\ldots$$



Two cases

Now there are Bayesians who have developed rules for obtaining reference prior distributions that purport to express little or no information regarding the parameter θ. All of these methods, except Geisser's and Zellner's, yield the same reference priors

$$P_B(\theta)\propto\theta^{-1/2}(1-\theta)^{-1/2}$$

for the binomial and

$$P_N(\theta)\propto\theta^{-1}(1-\theta)^{-1/2}$$

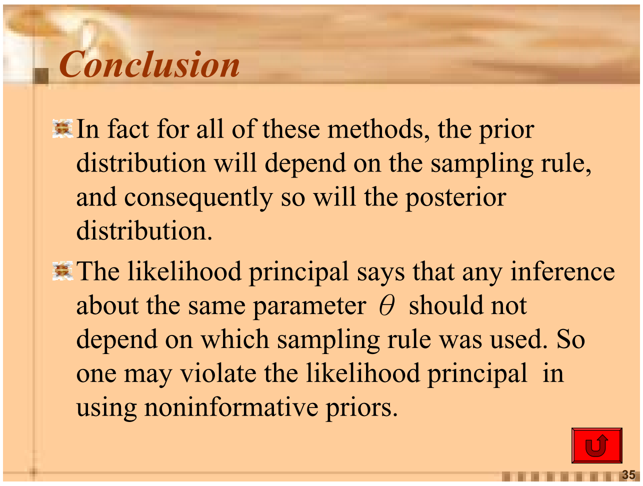
for the negative binomial case. Hence the posterior densities for these two cases are

$$P_B(\theta\mid t,n)\propto\theta^{t-1/2}(1-\theta)^{n-1-1/2}$$

and

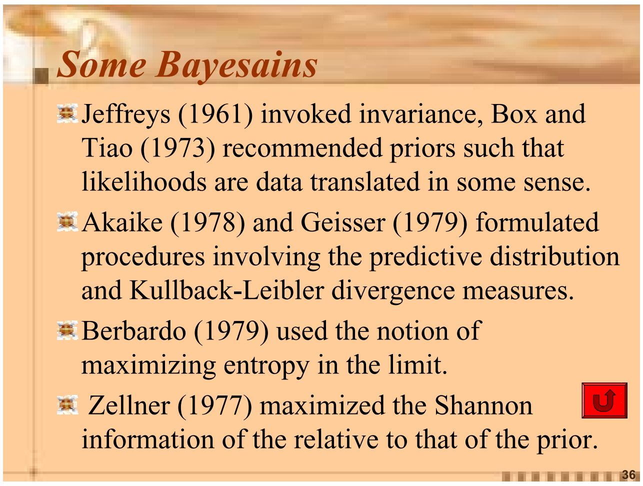
$$P_N(\theta\mid t,n)\propto\theta^{t-1}(1-\theta)^{n-t-1/2}$$

respectively.



Conclusion

- In fact for all of these methods, the prior distribution will depend on the sampling rule, and consequently so will the posterior distribution.
- The likelihood principal says that any inference about the same parameter θ should not depend on which sampling rule was used. So one may violate the likelihood principal in using noninformative priors.



Some Bayesains

- Jeffreys (1961) invoked invariance, Box and Tiao (1973) recommended priors such that likelihoods are data translated in some sense.
- Akaike (1978) and Geisser (1979) formulated procedures involving the predictive distribution and Kullback-Leibler divergence measures.
- Berbardo (1979) used the notion of maximizing entropy in the limit.
- Zellner (1977) maximized the Shannon information of the relative to that of the prior.