Baysian Statistics

11/05/23

In Statistical Inference, decisions about populations, such as what are the mean or variance of some characteristic are based on sample data.

Statistical inference can therefore be regarded as a game between Nature, which controls the relevant features of a population and the Statistician, who is trying to make a decision about the population.

In Statistical games, the Statistician doesn't know the nature's strategy but may have some information about the same through sample data

Q.1] A statistician is told that a coin has either a head on one side and a tail on the other or it has two heads. The statistician cannot inspect the coin but can observe a single toss of the coin and see whether it shows a head or tail. The statistician must then decide whether or not the coin is two headed. If the Statistician makes the wrong decision there is penalty of ₹ $1. There is no penalty (or reward) if right decision is made.

Ignoring the fact that the statistician can observe ~~the~~ one toss the problem could be regarded as following game

Statistician (Player A)

Nature (Player B)	a_1	a_2
θ_1	0	1
θ_2	1	0

where $\theta_1 \Rightarrow$ "the state of nature" is that the coin is two headed

$\theta_2 \Rightarrow$ "the state of nature" is that the coin is balanced.

$a_1 \Rightarrow$ Statistician's decision is that the coin is two headed.

$a_2 \Rightarrow$ Statistician's decision is that coin is balanced.

However, statistician knows ~~that~~ what has happened on the toss of the coin, ie. statistician knows whether a random variable x (say) has taken the value $x=0$ (heads) or $x=1$ (tail). The statistician will want to use this information in choosing b/w a_1 and a_2 and needs a "decision function" setting out the action to take in each case.

One possible decision function would be decision function $d_1(x) = \begin{cases} a_1 ; x=0 \\ a_2 ; x=1 \end{cases}$

other decision function might be $d_2(x) = \begin{cases} a_1 ; x=0 \\ a_1 ; x=1 \end{cases}$ | $d_3(x) = \begin{cases} a_2 ; x=0 \\ a_2 ; x=1 \end{cases}$

$$d_4(x) = \begin{cases} a_2 ; x=0 \\ a_1 ; x=1 \end{cases}$$

$d_1(0)=a_1$ $d_1(1)=a_2$	$d_2(0)=a_1$ $d_2(1)=a_1$	$d_3(0)=a_2$ $d_3(1)=a_2$	$d_4(0)=a_2$ $d_4(1)\ a_1$

These are only possible decision functions if the statistician's value is based purely on the observed value x and randomized choices are not allowed.

Some of these may not be sensible in practice.

The entries in table give corresponding values of the loss function thus:

Nature \ Statistician	a_1	a_2
θ_1	$L(a_1,\theta_1)$	$L(a_2,\theta_1)$
θ_2	$L(a_1,\theta_2)$	$L(a_2,\theta_2)$

Now consider our first decision function i.e. $d_1(x)$ and calculate expected resulted loss.

$$R(d_i,\theta_j) = E\left[L(d_i(x),\theta_j)\right]$$; where $L(d_1(x),\theta_j)$ is loss when θ_j is nature's strategy and d_i is decision taken by statistician.

R gives the value of expected loss from particular decision function and state of nature.

The expectation is taken with respect to random variable x and under θ

$R(d_1,\theta_1)$	under θ_1 2 headed coin	$P(x=0)=1$ $P(x=1)=0$
$R(d_1,\theta_2)$	under θ_2 (1H, 1T) unbiased coin	$P(x=0)=1/2$ $P(x=1)=1/2$

$$R(d_1,\theta_1) = L(d_1(x),\theta_1)$$
$$= 1 \times L(d_1(0),\theta_1) + 0 \times L(d_1(1),\theta_1)$$
$$= 1 \times L(a_1,\theta_1) + 0 \times L(a_2,\theta_1)$$
$$= 1 \times 0 + 0 \times 1$$
$$R(d_1,\theta_1) = 0$$

$$R(d_1,\theta_2) = \frac{1}{2}\times L(a_1,\theta_2) + \frac{1}{2}\times L(a_2,\theta_2) = \frac{1}{2}\times 1 + \frac{1}{2}\times 0 = \frac{1}{2}$$

$$R(d_2,\theta_1) = 1\times L(a_1,\theta_1) + 0\times L(a_2,\theta_1) = 1\times 0 + 0\times 1 = 0$$

$$R(d_2,\theta_2) = \frac{1}{2}\times L(a_1,\theta_2) + \frac{1}{2}\times L(a_2,\theta_2) = \frac{1}{2}\times 1 + \frac{1}{2}\times 1 = 1$$

$$R(d_3,\theta_1) = 1\times L(a_2,\theta_1) + 0\times L(a_2,\theta_1) = 1\times 1 + 0\times 1 = 1$$

$$R(d_3,\theta_2) = \frac{1}{2}\times L(a_2,\theta_2) + \frac{0}{2}\times L(a_2,\theta_2) = \frac{1}{2}\times 0 + \frac{1}{2}\times 0 = 0$$

$$R(d_4,\theta_1) = 1\times L(a_2,\theta_1) + 0\times L(a_1,\theta_1) = 1\times 1 + 0\times 0 = 1$$

$$R(d_4,\theta_2) = \frac{1}{2}\times L(a_2,\theta_2) + \frac{1}{2}\times L(a_2,\theta_2) = \frac{1}{2}\times 0 + \frac{1}{2}\times 1 = \frac{1}{2}$$

Statistician

Nature	d_1	d_2	d_3	d_4
θ_1	0	0	1	1
θ_2	1/2	1	0	1/2

This matrix shows the expected loss for different decision rules. rather then the actual payoffs for different courses of action.

By inspection it can be seen that d_1 dominates d_2 and d_3 dominates d_4 so d_2 and d_4 can be discarded (they are said to inadmissible). This is not surprising as they imply that a_1 (that coin is two headed is accepted even though a tail was observed.

The game thus reduces to 2×2 zero sum two person game.

Statistician

Nature	d_1	d_3
θ_1 → P	0	1
θ_2 → 1-P	1/2	0

$$E(R(d_1,\theta_1)) = 0\times p + \frac{1}{2}(1-p)$$
$$E(R(d_3,\theta_1)) = 1\times p + (1-p)0$$

$$\rightarrow \frac{1}{2}(1-p) = p$$

$$p = 2/3$$

$$1-p = 1/3$$

Selection of decision after ruling out higher loss decision

① minmax criteria

② Bayes Theorem

when nature_state strategy is θ_1

$$E(\underset{\text{Loss}}{R}, \theta_1) = 0 \times p + (1-p) \times 1 = 1-p$$

when θ_2,

$$E(R, \theta_2) = \frac{1}{2} \times p + 0 \times (1-p) = p/2$$

$$\Rightarrow 1-p = p/2$$

$$\Rightarrow p = 2/3$$

& the randomized strategy that minimizes the maximum loss is choose d_1 two thirds and d_3 one thirds of time; $1/3$ is value of game & expected loss function when $p = 2/3$

If $p > \frac{1}{3}$ the statistician will take d_1; if $p < \frac{1}{3}$ then d_4.

when $p = \frac{1}{3}$, two Bayes risk are equal & either d_1 or d_4 can be choosen

$\frac{1}{3}$ is value of game also known as expected risk (expected value risk)

Q-2.] A statistician is observing values from Bin $(2, p)$ distribution. He knows that p is equal to either to $\frac{1}{4}$ or $\frac{1}{2}$ and he is trying to choose between these two values. He observes a single value x from the distribution. He proposes to use one of the following four decision functions:

$d_1(x)$: Set $p = 1/4$ when $x = 0$
Set $p = 1/2$ when $x = 1$ or 2

$d_2(x)$: set $p = 1/4$ when $x = 0$ or 1
set $p = 1/2$ when $x = 2$

$d_3(x)$: set $p = \frac{1}{4}$ when $x = 0, 1$ or 2

$d_4(x)$: set $p = 1/2$ when $x = 0, 1$, or 2

If he incorrectly concludes that $p = 1/4$ he suffers a loss of 1. If he incorrectly concludes that $p = 1/2$, he suffers a loss of 2.

Find the risk function for each decision function and find the decision function that minimizes the maximum expected loss.

Solution : Consider first decision function $d_1(x)$. If p actually is $1/4$, there is a loss of 0 if $x=0$ and loss of 2 if $x=1$ or 2.

Nature \ Statistician	$1/4$	$\frac{1}{2}$
$1/4$	0	2
$1/2$	1	0

$$R(d_1, p=1/4) = 0 \times P(x=0|p=\tfrac{1}{4}) + 2 \times P(x=1 \text{ or } 2|p=\tfrac{1}{4})$$
$$= 2 \times P(x=1 \text{ or } 2|p=\tfrac{1}{4})$$

$$R(d_1, p=\tfrac{1}{4}) = 2 \times 7/16 = 7/8$$

$$R(d_1, \theta_1) = R(d_1(x), \theta_1)$$
$$= E[L(d_1(x)), \theta_1]$$

$$P(x=x) = {}^2C_x\, p^x (1-p)^{2-x}$$

$$P[x=0] = {}^2C_0\, p^0 (1-p)^2$$
$$= {}^2C_0 \left(\tfrac{3}{4}\right)^2 \quad \because p = \tfrac{1}{4}$$

$$P(x=0) = \frac{9}{16}$$

$$P[x=1 \text{ or } 2] = 1 - P(x=0)$$

$P[x=1 \text{ or } 2] = 7/16$ here $p=\frac{1}{4}$

$$P[x=2] = {}^2C_2 \left(\tfrac{1}{2}\right)^2 \left(\tfrac{1}{2}\right)^0 \quad \because p = \tfrac{1}{2}$$

$$P[x=2] = \tfrac{1}{4}$$

$$P(x=0) = {}^2C_0 \left(\tfrac{1}{2}\right)^0 \left(\tfrac{1}{2}\right)^2$$

$$P(x=0) = 1/4$$

$$P(x=2|p=\tfrac{1}{4}) = {}^2C_2 \left(\tfrac{1}{4}\right)^2 \left(\tfrac{1}{4}\right)^0$$
$$= \frac{1}{16}$$

$$P[x=0 \text{ or } 1] = 1 - P(x=2)$$
$$= 1 - \tfrac{1}{4}$$
$$P[x=0 \text{ or } 1] = \frac{3}{4}$$

Similarly

$$R(d_1|p=\tfrac{1}{2}) = 1 \times P(x=0|p=\tfrac{1}{2}) + 0 \times P(x=1 \text{ or } 2|p=\tfrac{1}{2})$$

$$R(d_1|p=\tfrac{1}{2}) = 1 \times \tfrac{1}{4}$$

Now, repeat for other risk functions.

$$R(d_2, p=\tfrac{1}{4}) = 0 \times P(x=0 \text{ or } 1|p=\tfrac{1}{4}) + 2 \times P(x=2|p=\tfrac{1}{4})$$

$$R(d_2, p=\tfrac{1}{4}) = 2 \times \frac{1}{16} = \frac{1}{8}$$

$$R(d_2, p=\tfrac{1}{2}) = 1 \times P(x=0 \text{ or } 1|p=\tfrac{1}{2}) + 0 \times P(x=2|p=\tfrac{1}{2})$$

$$R(d_2, p=\tfrac{1}{2}) = 1 \times \frac{3}{4} = \frac{3}{4}$$

~~$R(d_3, p=\frac{1}{4}) =$~~

For d_3 we always choose $p=\frac{1}{4}$. So there ~~is~~ will be a loss of 0 if p actually is $\frac{1}{4}$ and a loss of 1 if $p=\frac{1}{2}$.

$$R(d_3, p=\tfrac{1}{4}) = 0$$

$$R(d_3, p=\tfrac{1}{2}) = 1$$

Similarly for d_4

$$R(d_4, p=\tfrac{1}{4}) = 2$$

$$R(d_4, p=\tfrac{1}{2}) = 0$$

nature's strategy ↓	d_1	d_2	d_3	d_4
$p = \frac{1}{4}$	$7/8$	$1/8$	0	2
$p = \frac{1}{2}$	$1/2$	$3/4$	1	0

If we look at the risk functions we find that the maximum loss under each of the function is $7/8$ for d_1, $3/4$ for d_2, 1 for d_3 and 2 for d_4.

Note: first circle maximum value in columns i.e. select decision itself then compare minimum is choosed

So decision function that minimizes the maximum expected loss is d_2.

* Decision Criteria: In general, it is possible to find the best decision function only in respect of some criteria. Two criteria will be considered here.

① * The minimax criteria :-

Under the minimax criteria, the decision function d chosen is that for which $R(d,\theta)$, maximised with respect to θ is minimum.

Note: here that we are now applying the minimax criteria to whole strategy, rather than to the choices on each item.

Applying the minimax criteria to example (Q.1) with d_2 and d_4 discarded, the maximum risk for d_1 is $\frac{1}{2}$ and for d_3 is 1 and so d_1 minimizes the maximum risk

② The Bayes Criteria: If (H) is regarded as random variable under the Bayes criterion the decision function choosen is that for which $E[R(d,H)]$ is minimum where the expectation is taken with respect to (H). The criterion needs (H) to be regarded as random variable with a given distribution.

Applying the Bayes criterion to example (Q.1) requires probabilities to be attached to Nature's two strategies θ_1 and θ_2. If $P(\theta_1) = p$ and $P(\theta_2) = 1-p$

then bayes risk for d_1 is $0 \cdot p + \frac{1}{2}(1-p) = \frac{1}{2}(1-p)$

d_3 is $1 \cdot p + 0(1-p) = p$

$p = \frac{1}{2}(1-p)$

$p = \frac{1}{3}$

Thus when $p > \frac{1}{3}$ the Bayes risk of d_1 is less than the Bayes risk of d_2 and d_1 is preferred to d_3

When $p < \frac{1}{3}$ the Bayes risk of d_3 is less than the Bayes risk of d_1 and d_3 is preferred to d_1

When $p = \frac{1}{3}$ the two Bayes risks are equal and either d_1 or d_3 can be chosen.

— x —

For Q.2.

If statistician has prior feeling it is about equally likely that p will be equal to $\frac{1}{4}$ or $\frac{1}{2}$

then

$E(R(d_1, (H))) = \frac{7}{8} \times p + (1-p)\frac{1}{2} = \frac{7p}{8} + \frac{1}{2} - \frac{p}{2} = \frac{3p}{8} + \frac{1}{2}$

$E(R(d_2, (H))) = \frac{1}{8} \times p + (1-p)\frac{3}{4} = \frac{p}{8} + \frac{3}{4} - \frac{3p}{4} = \frac{3}{4} - \frac{5p}{8}$

$E(R(d_3, (H))) = 0 \times p + (1-p)1 = 1-p$

$E(R(d_4, (H))) = 2 \times p + (1-p)0 = 2p$

$1 - p = 2p$

$\boxed{p = \frac{1}{3}}$

13/05/23

Likelihood is occurrence of $f(x,\theta)$. It is function of $x_1, x_2 \dots x_n$

pdf of only parameter

$x_i \sim f(x_i,\theta)$

$L(\theta|x) = L(\theta) = \prod_{i=1}^{n} f(x_i,\theta)$

prior $\leftarrow$ $p(x) = g(\theta)$

$$p(\theta|x) = \frac{L(\theta).g(\theta)}{\int L(\theta).g(\theta)\,d\theta}$$

$L(\theta).g(\theta) \Rightarrow$ joint

$\int L(\theta).g(\theta)\,d\theta \Rightarrow$ marginal

$p(\theta|x)$: posterior.

$$p(\theta|x) \propto L(\theta).g(\theta)$$

proper prior - integration of proper prior is 1

improper prior - integration of improper prior not 1 but ∞

Q.1 $L(\theta) \sim Poi(\lambda)$

~~$L(\theta) = \prod_{i=1}^{n} e^{-\lambda}\lambda^{x}$~~

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n}(x!)}$$

$g(\lambda) = Exp(\lambda')$

$= \lambda' e^{-\lambda\lambda'}$

~~$p(\theta|\lambda) = L(\theta).g(\lambda)$~~

$$p(\lambda|x) = \frac{L(\lambda)\,g(\lambda)}{\int L(\lambda)\,g(\lambda)\,d\lambda}$$

* Prior and Posterior Density -

Let $x_1, x_2 \dots x_n$ be random variable sample from a population specified by the density or population function $f(x;\theta)$ and it is required to estimate θ.

In Baye's paradigm, the parameter θ is considered as random variable, therefore it will have a ~~prop~~ probability distribution. This allows the use of any knowledge available about possible values of θ before collection of any data. This knowledge is quantified by expressing it as the prior distribution of θ. Then after collecting appropriate data the posterior distribution of θ is determined and thus forms the basis of all inferences concerning θ. This information from random sample is contained in the likelihood function for that sample. So, the Bayesian approach combines the information obtained from the likelihood function with the information in the prior distribution and hence obtained a posterior estimate for the required population parameter.

$$\rightarrow p(\theta|\underset{\sim}{x}) = \frac{L(\theta)\cdot g(\theta)}{\int L(\theta)\cdot g(\theta)\,d\theta} \Rightarrow p(\theta|\underset{\sim}{x}) \propto L(\theta)\cdot g(\theta)$$

where $\propto = 1$

If $L(\theta)$ is likelihood and $g(\theta)$ is prior

Pg No. 19 Bayesian CIMS CSA.pdf

Q-1] B_1 ⇒ Shorter life time Bulbs ⇒ mean life ⇒ 500 hrs ~ exp $(\frac{1}{500})$

B_2 ⇒ Longer " " " " " ⇒ 2500 hrs ~ exp $(\frac{1}{2500})$

approx 5 bulbs still going after ⇒ approx 300 hrs ⇒ 5 bulb alive after 300 hrs = A

$P(B_2 \mid 5 \text{ bulb alive after 300 hrs}) = ?$

$$P(A|B_1) = \left[\int_{300}^{\infty} \frac{1}{500} e^{-\frac{1}{150}x} \cdot dx\right]^5 = e^{-3} \quad \Big| \quad P(A|B_2) = e^{-3/5}$$

$P(B_1) = \frac{1}{2}$; $P(B_2) = \frac{1}{2}$.

$$P(B_1|A) = \frac{P(A|B_1)\,P(B_1)}{P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2)} \quad \Big| \quad P(B_2|A) = \frac{P(A|B_2)\,P(B_2)}{P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2)}$$

$$P(B_2|A) = \frac{e^{-3/5}\cdot \frac{1}{2}}{e^{-3}\cdot\frac{1}{2} + e^{-3/5}\cdot\frac{1}{2}} = \frac{e^{-1/5}}{1+e^{-1/5}}$$

Gamma distribution

$$G(\alpha,\lambda) = \frac{\lambda^{\alpha}}{\sqrt{\alpha}} \cdot e^{-\lambda x} x^{\alpha-1} = \frac{\lambda^{\alpha}}{\sqrt{\alpha}} \int_0^{\infty} e^{-\lambda x} \cdot x^{\alpha-1}\,dx$$

$$G(\alpha,1) = \frac{e^{-x}\cdot x^{\alpha-1}}{\sqrt{\alpha}}$$

$$G(1,\lambda) = \lambda e^{-\lambda x}$$

$G(\frac{1}{2}, \frac{n}{2})$ ⇒ (putting $\alpha = \frac{1}{2}$ and $\lambda = \frac{n}{2}$) this will give Chi-square distribution.

Baye's Risk

$$E_{\theta|x}[L(\theta)] = \int_{R_\theta} L(\theta).p(\theta|x).d\theta$$

R_θ → Range

$p(\theta|x)$ → posterior distribution

$$E_x(g(x)) = \int_{R_x} g(x).g(x)\,dx$$

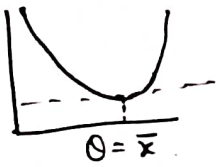
$= p(\theta|x)$

posterior distribution mean $= \int \theta.p(\theta|x)\,d\theta$

$$\Gamma(\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma\alpha}\int_0^\infty e^{-\lambda x}.x^{\alpha-1}.dx = 1$$

$$\int_0^\infty e^{-\lambda x}.x^{\alpha-1}.dx = \frac{\Gamma\alpha}{\lambda^\alpha}$$

① Quadratic error loss function

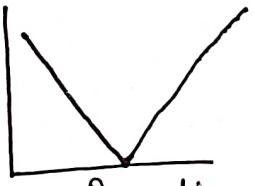


$E(g(x)-\theta)^2$

⇓ loss will minimum about mean

② Absolute error loss function

$|g(x)-\theta|$

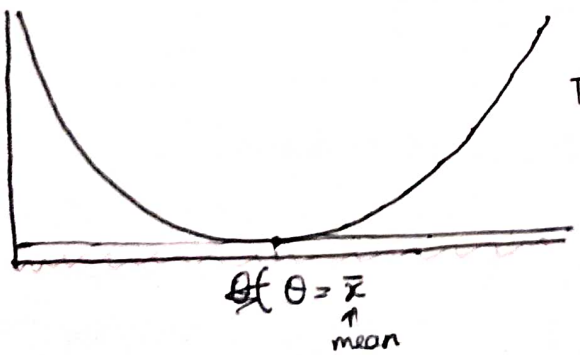


loss will be minimum about median

The Loss Function - To obtain an estimator of unknown parameter θ, a loss function must be specified. This is a measure of the "loss" incurred when $g(\underline{x})$ is used as an estimator of θ. A loss function is sought which is zero when the estimation is exactly correct, that is $g(\underline{x}) = \theta$ and the loss is zero and it is positive and does not decrease as $g(\underline{x})$ gets further away from θ. There is one very commonly used loss function called quadratic or squared error loss. Two others are also in practice.

① Quadratic error loss function (mean squared error loss function)

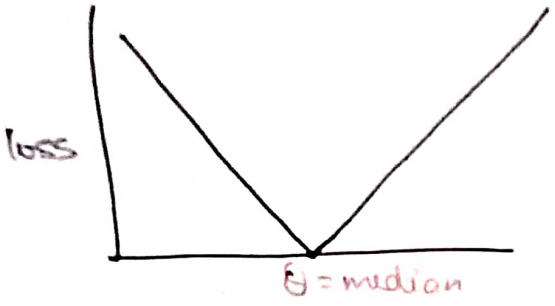
$$L(g(\underline{x}),\theta) = [g(\underline{x}) - \theta]^2$$



The squared error loss function is minimum when $g(x)$ is equal to the parameter value θ.

(2) Absolute error loss function.

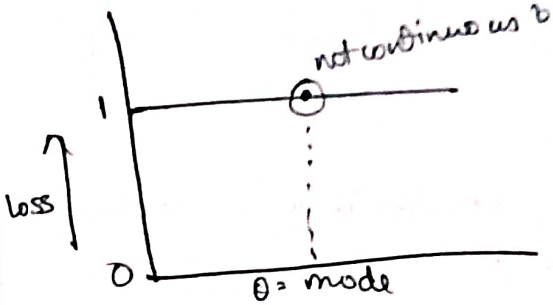
$$L(g(\underline{x}),\theta) = |g(\underline{x})-\theta|$$



loss will be minimum when θ = median.

(3) 0-1 Loss function (All or nothing loss)

$$L(g(\underline{x}),\theta) = \begin{cases} 0 & ; g(\underline{x}) = \theta \\ 1 & ; g(\underline{x}) \neq \theta \end{cases}$$



~~If $g(\theta)$ is not $g(\underline{x})$ is~~

If $g(\underline{x})$ is equal to θ then there is always loss

The Baysian ~~Bayes's~~ estimator that arises by minimizing the expected loss for each of these loss functions in turn is the mean, median and mode respectively of the posterior distribution, each of which is a measure of a location of the posterior distribution.

The expected posterior loss is

$$EPL = E\{L(g(\underline{x}),\theta)\} = \int L(g(\underline{x}),\theta).f(\theta|\underline{x}).d\theta$$

Q. 10 i.i.d observation from a Poisson(λ) distribution give 3, 4, 3, 1, 5, 5, 2, 3, 3, 2. assuming Exp(0.2) prior distribution for λ. find baysian estimation λ under squared error loss.

$X_i \sim Poi(\lambda)$

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad ; \lambda>0, \; x>0$$

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod_{i=1}^{n}(x!)}$$

$\lambda \sim Exp(\lambda')$

$$g(\lambda) = \lambda' e^{-\lambda\lambda'} \; ; \; \lambda, \lambda' > 0$$

$$P(\lambda|x) = \frac{(e^{-n\lambda}\lambda^{\Sigma x_i} \cdot \lambda' e^{-\lambda\lambda'})/\prod_{i=1}^{n}(x!)}{\int_0^\infty L(\lambda)\, g(\lambda)\, d\lambda}$$

$$P(\lambda|x) = \frac{e^{-(n+\lambda')\lambda} \cdot \lambda^{\Sigma x_i}}{\int_0^\infty e^{-(n+\lambda')\lambda} \cdot \lambda^{\Sigma x_i + 1 - 1} \cdot d\lambda} = \frac{e^{-(n+\lambda')\lambda} \cdot \lambda^{\Sigma x_i}}{\lfloor \Sigma x_i + 1 \rfloor / (n+\lambda')^{\Sigma x_i + 1}}$$

$\lambda' = 0.2$, $n = 10$

$\Sigma x_i = 3+4+3+1+5+5+2+3+3+2$

$\Sigma x_i = 31$

$$P(\lambda|x) = \frac{e^{-(10+0.2)\lambda} \cdot \lambda^{31}}{\lfloor 31+1 \rfloor / (10+0.2)^{32}} = \frac{(10.2)^{32} \cdot e^{-10.2\lambda} \cdot \lambda^{31}}{\lfloor 32}$$

25/05/23 Upadhyay Sir

Sample space	n(event)	Probability
10	4	4/10
100	52	52/100
1000	497	497/1000
⋮	⋮	⋮

after some time, it gives some constant value for $n \to \infty$

$$P(A) = \lim_{n\to\infty} \left(\frac{n}{m}\right) = c \leftarrow \text{constant}$$

($\frac{n}{m}$: probability)

Decision Theory

In decision theory, we mainly talk about losses

↑ Negative losses.

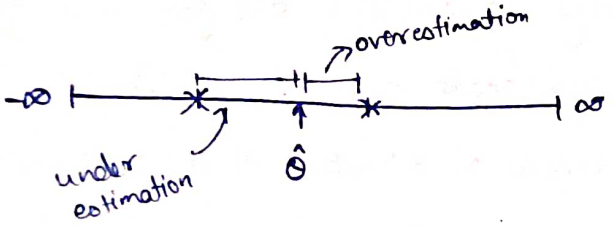
eg. [New company] → loss / gain

Decision maker

Without any past information or experience, we can't open new company (prior)

Domain knowledge

Who defines loss mathematically



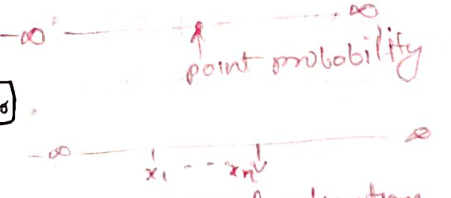
$$\frac{|\theta - \hat{\theta}|}{(\theta - \hat{\theta})^2}$$

Statistics is non ~~deterministic~~ deterministic in probabilistic sense.

Always deals with uncertainty.

$X \sim f(x|\theta)$

- To draw inference about parameter θ
- parametric statistical inference
- parameter: something which is fixed but unknown
- $\theta \in \mathbb{R}$
- $\theta \in \mathbb{R}$ or R^+ or R^- or $(a,b) \in R$



interval estimation

- θ_0's correctness through hypothesis.
- Basis of what we draw all inferences is Random sample (x_i)
- Likelihood f^n packs entire information given by $x_1, x_2 \cdots x_n$

$L(\theta, \underset{\sim}{x})$ or $L(\underset{\sim}{x}, \theta)$

Statistician work

⇓

to minimize the cost loss

Both begin with sample distribution

loss of average is called risk

Classical or frequentist

$X \sim f(x|\theta)$

$L(\underset{\sim}{x}, \theta)$

Likelihood function is used as basis of inference

⇒ using inductive logic ⇒ indirect drawing of inference.

we have one set of information $L(x, \theta)$

Baye's Inference

$\theta \sim f(x|\theta)$

$\theta \sim \pi(\theta)$ → distribution of parameters (prior distribution)

If θ is fixed, how come it is Random variable.

$x_1 \cdots x_n$ - random sample.

$L(\underset{\sim}{x}, \theta)$ → likelihood fn

$\theta \sim p(\theta|\underset{\sim}{x})$ ⇒ posterior dist..

⇒ Combining Prior & Likelihood fn to form Posterior.

⇓

basis of baysian Inference, logic is deductive (direct) of drawing inference

we draw conclusion on basis of that we have two sets of information i.e. $\theta \sim \pi(\theta)$ & $L(\underset{\sim}{x}, \theta)$

Bayesian Inference considers $\bar{X}$ as fixed from sample $\therefore$ Sample mean goes fixed.

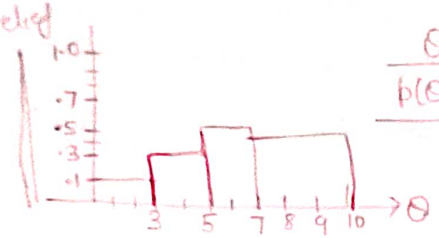
Baye's Theorem updates your prior belief based on likelihood fn.

θ is regarded as Random Variable is our belief.

→ Bayesian Statistics ⇒ science of changing or updating our beliefs.
→ we have belief combined with belief obtained through observation then we update our beliefs.

$\theta \in (0, 10)$

belief



θ :	0-3	3-5	5-7	7-10
$p(\theta)$:	0.1	0.3	0.5	0.4

↑ probability of belief is prior distribution.

prior elicitation ⇒ Top belief ~~we~~ that we have and give it a mathematical formulation to it ie. give it a probabilistic Model. It is not easy for some such as there are cases where we can't divide our belief on height of 5.2 and 5.3 we will get confused we can not tell sweetness of mango without tasting it, so we cannot draw our belief.

Belief can be turned into probabilistic model by defining it in mathematical numbers such as from $\underset{\text{min}}{0}$ to $\underset{\text{max}}{1}$ or $\underset{\text{min}}{0}$ to $\underset{\text{max}}{1000}$

values that belief can have.

Subjective probability + utility

Axioms

3 axioms ⇒ ① non-negativity
② Sigma (Additivity)
③ Probability of whole space is one
④ non-measure

* Consistency
* Coherence

* Utility ⇒ negativity of loss
* loss on average ⇒ Risk
* Diff b/w Bayesian & classical paradigm.

* Drag the belief in mathematical form is Prior elicitation.

① Histogram Approach ⇒

Histogram approach is method of determining the prior density of parameter.

(i) works by dividing data into a number of intervals and assign a subjective probability to each (x).

(ii) method of determining prior density of parameter $\in \boxed{\theta \in R}$

(iii) used to determine the relative likelihood of different values of parameter.

Example ⇒ Data is normally distributed then the histogram will be bell shaped

(i) the most likely value of the parameter will be mean of the distribution, least likely value will be the tails of the distribution. The relative likelihood of other values of the parameter can be determined by comparing the height of histogram.

② Relative likelihood Approach ⇒

this approach is method to determining a prior density based on our subjective beliefs about the relative likelihood of different value of the parameter.

(i) we first identify the value of parameter that we have most belief in ⇒ then
- assign the higher subjective probability to these values.
- ⇒ assign the lower subjective probability to the values of the parameter that we believe are less likely.

(ii) The relative likelihood approach can be used to represent a prior density even if we do not have any specific information about the parameter.

example ⇒ let's say we are interested in the probability as that coin is biased towards head.

⇒ we could use the relative likelihood approach to represent our prior density about the probability of head.

⇒ based on our belief we most likely 0.5, 0.6, 0.7.

Advantages ⇒ (1) Simple & easy to understand (2) It can be used to represent our subjective belief about the ___ (3) It can be used even if we do not have any specific information about the parameter.

Disadvantages ⇒ the result of the approach will depend on our own personal beliefs.
⇒ approach can be subjective and difficult to justify.

(3)

Matching given functional form ⇒

(i) the matching given functional ~~form~~ form approach to subjective determination of the apr prior density is a method of determining a prior density by matching it ~~is~~ to function for.

→ It can be used to represent our ~~ob~~ subject beliefs about the parameter using a specific functional form.

→ It can be used even if we do not have any specific information about the parameter.

↳ some example as functional forms that can be used in the matching given for ⇒

- Beta — $B(\alpha,\beta)$
- Gamma — $G(\alpha,\beta)$
- normal — $N(\mu,\sigma^2)$
- uniform.

→ Hyper-parameters ⇒ (i) Ascertain these parameters (hyperparameters)
(ii) Determine the shape of distribution from an expert (family of distribution)

(iii)

(4)

*CDF ~~Distribution~~ Determination : $F_{(H)}(\theta) = P[(H) \le \theta]$

Multivariate Prior Elicitation

$x \sim f(x|\theta)$; $\theta = \theta_1 \dots \theta_k$; $k \ge 2$

$x \sim f(x|\theta_1,\theta_2)$ | $\pi(\theta_1,\theta_2)$

Bivariate — visualize joint probability of θ_1 & θ_2

→ Normally, Conditional Probability we study in 1D.

Priori Independent : $\pi(\theta_1,\theta_2) = \pi(\theta_1).\pi(\theta_2)$

$\pi(\theta_1,\theta_2 \dots \theta_k) = \pi(\theta_1).\pi(\theta_2)\dots\pi(\theta_k)$

$\pi(\theta_1,\theta_2) = \pi(\theta_1).\pi(\theta_2|\theta_1)$

marginal distribution

Condition Probability is Bivariate (since 2 variables A & B)

08|07|23

Reema Mam

Classical ← here we only have info about sample

Baysian ← here we have info about sample and prior distribution ↓ about θ.

Statistical Inference

Observed (observation or sample values)

Unobserved (Parameter) — fixed but unknown ↔ population characteristics

$X \sim f(x_i, \theta)$ ↓ represents population

Distribution is given in both classical & baysian

Check ✓ ~~prior distribution gives info about the parameter θ~~

⇒ Prior Information is all about the parameter, not about the sample.

Baysian ⟶ Prior (↓ about θ (parameter)) and Sample Values ⟶ updated information is [Posterior] ~~posterior~~

* If prior distribution $g(\theta, \eta)$ is improper then posterior may or may not be improper. $[\hat{p}(\theta|x)]$

$X \sim f(x, \theta)$

$\theta \sim g(\theta, \eta)$

$p(\theta|x) \propto L(x, \theta) \cdot g(\theta, \eta)$

$p(\theta|x)$: posterior fn

$L(x, \theta)$: likelihood is fn of θ (parameter)

we are treating θ as random variable

Types of Prior:

Broadly classified into ① Informative Prior ② Non Informative Prior

(a) If weak prior knowledge (we don't know info about the parameter.) ⇒ Non-informative

(b) If Strong prior knowledge ⇒ Informative about the parameter

$\theta \sim g(\theta, \eta)$ or $g(\eta)$

Proper Prior ↓ $\int g(\theta)\, d\theta = 1$

Improper Prior $\int g(\theta)\, d\theta = \infty$

$X \in (-\infty, \infty)$ → $\int_{-\infty}^{\infty} f(x, \theta)\, dx$

$X \to \mathcal{X}$ (Sample Space)

$\theta \to \Theta$ (Parameter Space)

Type of parameter

① location parameter (+, −) ie. Shift of origin | ② scale parameter (*, /) Scale parameter = $\frac{1}{\text{rate}}$ | ③ Shape parameter (changing shape of curve) or it changes increasing/decreasing rates

Example of scale & location parameter

$X \sim N(\mu, \sigma^2)$

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); \quad -\infty < x, \mu < \infty, \ \sigma > 0$$

$\sigma^2 \Rightarrow$ scale parameter

$\mu \Rightarrow$ location parameter.

Non-Informative Prior Distribution

A prior which contains no-information about the parameter (say θ) or more crudely "which favours" no possible values of θ over other maybe called non-informative prior

For example ⇒ In testing two simple hypothesis, the prior which gives probability $\frac{1}{2}$ to each hypothesis is clearly non-informative

$N(\theta, 1)$; $\theta \in (-\infty, \infty)$ } Suppose the parameter of interest is a normal mean θ, so that parameter space is $\theta \in (-\infty, \infty)$. If a non-informative prior density is desired, it seems reasonable to give equal weight to all possible values of θ. Unfortunately, if $\pi(\theta) = c > 0$ is chosen, then π has infinite mass (ie. $\int \pi(\theta) d\theta = \infty$) and is not proper density. Nevertheless, such π can be successfully worked with. The choice of c is unimportant, so that typically the non-informative prior density for this problem is chosen to be $\pi(\theta) = 1$. This is often called uniform density on R^1.

$\pi(\theta) \propto c$ or $\pi(\theta) = c > 0$

$\pi(\theta) = 1$

Note
It will frequently happen that natural non-informative prior is an improper prior. Namely if π has infinite mass.

15/07/23 * Determination of non-informative prior.

Case 1: (1) When 'Θ' is finite: Suppose (H) i.e. parameter space is finite and it contains 'n' elements. Θ: 1, 2 . . . n.

The obvious non-informative prior is to then give each element of (H) probability 1/n.

(H): $\theta_1, \theta_2 \cdots \theta_n$

$g(\theta)$: $\frac{1}{n} \cdot \frac{1}{n} \cdots \frac{1}{n}$

↑ H is proper but if $n \Rightarrow \infty$ then it is improper

One might generalize it by considering ~~infinite (H)~~ $n = \infty$ to a improper prior. A generalization $g(\theta)$ to infinite Θ may be proportional to a constant $[\forall \theta \in (H)]$ Consider a non-informative prior for parameter Θ then

$$\pi(\theta) \propto c$$

Instead of considering a Θ, suppose the problem has been parametrized in terms of $\eta = \exp(\theta)$. This is one-to-one transformation and so should have no bearing on the ultimate answer.

$\therefore \pi(\eta) \propto$ constant ⟶ location parameter.

But if $\pi(\theta)$ is the density for Θ then the corresponding density of η is

$$\pi(\eta) = \pi(\theta) \frac{d\theta}{d\eta} \quad \Rightarrow \text{Jacobian}$$

$$\pi(\eta) = \pi(\theta) \cdot \frac{1}{\eta}$$

$$\boxed{\pi(\eta) = \frac{\pi(\log \eta)}{\eta}}$$

$\eta = \exp(\theta)$

$\log \eta = \theta$

$\frac{1}{\eta} = \frac{d\theta}{d\eta}$

This is called lack of invariance of transformation. Hence, if the non-informative prior ~~(θ) is~~ for Θ is choosen to be constant, we should choose the non-informative prior for η to be proportional to $\frac{1}{\eta}$ to maintain consistency (and arrive at same answer in either parametrization). Thus, we cannot maintain consistency and choose both the non-informative prior for Θ and η that for η to be constant.

It could perhaps be argued that one usually chooses the most ~~intuitively~~ intuitively reasonable parametrization and that a lack of prior information should correspond to a constant ~~density~~ density in this parametrization, but the argument would be hard to defend in general.

The lack of invariance of constant prior has led to a search for non-informative priors which are appropriately invariant under transformation.

---

* Non-Informative Priors for Location and Scale Problems

① **Location Invariant Prior** ⇒ Suppose $\mathcal{X}$ (Sample Space) & $(\mathcal{H})$ (parameter space) both real and x has pdf of form $f(x-\theta)$ [ie. depends only on $x-\theta$]

The density is said to be a location density and θ is then called location parameter [or sometimes location vector when $p \geq 2$]

For example; $N(\theta, 1)$ and $\exp(\theta, 1)$ where $\theta \in (-\infty, \infty)$ is - location parameter

$N(\theta, \sigma^2)$ — θ: location parameter, σ^2: fixed

To derive the non-informative prior this situation, imagine that instead of observing X, we observe the random variable $Y = X + c$ $(c \in R^+)$.

Defining $\eta = \theta + c$, it is clear Y has density $f(y-\eta)$.

If now $\mathcal{X}$ & $(\mathcal{H})$ are R^+, then the sample space & parameter space for (Y, η) problem also R^+. The (X, θ) and (Y, η) problems are identical in structure and it seems reasonable to insist that they have the same non-informative prior.

Another of way of thinking this is to note that observing Y really amounts to observing X with different unit of measurement, one in which "origin" is c and not zero. Since, the choice of an origin for a unit of measurement is quite arbitrary, the non-informative prior should ~~peep~~ perhaps be independent of choice.

let π and π^* denote the noninformative priors in the (X,θ) and (Y,η) problems, respectively. We may assume that π and π^* are equal for any real space A.

So we can write $P^{\pi}(\theta \in A) = P^{\pi^*}(\eta \in A)$ ——— ①

for any set A in R^+.

Since $\eta = \theta + c$, it should also be true [by a simple change of variables] that

$$P^{\pi^*}(\eta \in A) = P^{\pi}((\theta+c) \in A)$$
$$P^{\pi^*}(\eta \in A) = P^{\pi}(\theta \in A-c) \quad \text{——— ②}$$
where $A-c = \{z-c\ ;\ z \in A\}$

from ① & ②

$$P^{\pi}(\theta \in A) = P^{\pi}(\theta\ A-c) \quad \text{——— ③}$$

Furthermore, this argument applies no matter what which $c \in R^+$ is chosen so that should hold for all $c \in R^+$. Any π satisfying this relationship is said to be location invariant prior.

Assuming that the prior has a density, we can write ③ as

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

If this hold for all sets A, it can be shown that it must be true that $\pi(\theta) = \pi(\theta-c)$ for all θ.

Setting $\theta = c$ thus gives $\pi(c) = \pi(0)$.

Recall, however, that this should hold for all $c \in R^+$. The conclusion is that π must be a constant function. It is convenient to choose the constant be 1, so the non-informative prior density for a location parameter is $\pi(\theta) = 1$. (This conclusion can be shown to follow from ③, even without the assumption that prior has density)

(scale parameter)

(2) Scale Invariant Prior ⇒ If A (one-dimensional) scale density is a density of the form $X \sim f(x;\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$; $\sigma > 0$

where X is random variable which having pdf $f(x,\sigma)$ or $f(x|\sigma)$

The parameter σ is called scale parameter.

Also, a sample of independent identically distributed random variables is said to be from a scale density if their common density is ~~scale~~ scale density

For example, $N(\mu, \sigma^2)$ and $\exp(\theta)$, parameter σ & θ are scale parameter of Normal & exponential distribution

To desire a non-informative prior for this situation, imagine that, instead of observing X, we observe the random variable $Y = cX$ $(c>0)$.

Defining $\eta = c\sigma$, an easy calculation shows that the density of Y is $\frac{1}{\eta} f\left(\frac{y}{\eta}\right)$.

$H \in (0,\infty)$

If now $\mathcal{X} = R^1$ or $\mathcal{X} = (0,\infty)$, then the sample ~~space~~ and parameter spaces for the (X,σ) problem are the same as those for the (Y,η) problem. The two problems are ~~fairly~~ thus identical in nature, or structure which again indicates that they should have some non-informative prior. [Here the transformation can be thought of as simply a change in the scale of measurement, from say inches to feet.]

Let π and π^* denote the priors in the (X,σ) and (Y,η) problems, respectively, this means that the equality

$$P^{\pi}(\sigma \in A) = P^{\pi^*}(\eta \in A) \quad \text{——— (1)}$$

should hold for all $A \in (0,\infty)$. Since $\eta = c\sigma$, it should also be true that

$$P^{\pi^*}(\eta \in A) = P^{\pi}(\sigma \in A/c) \quad \text{——— (2)} \quad \forall c > 0$$

where $(A/c) = \{c^{-1}z : z \in A\}$

Putting (1) & (2) together ⇒, it ~~should~~ follows that π should satisfy

$$P^{\pi}(\sigma \in A) = P^{\pi}(\sigma \in A/c) \quad \text{——— (3)}$$

This should hold for all $c>0$, and any distribution π for which this is true is called scale invariant.

The mathematical analysis of (8) proceeds as in the preceding example.

$$\int_A \pi(\sigma)\,d\sigma = \int_{c^{-1}A} \pi(\sigma)\,d\sigma = \int_A \pi(c^{-1}\sigma)\,c^{-1}\,d\sigma$$

$$\int_A \pi(\sigma)\,d\sigma = \frac{1}{c}\int_A \pi\left(\frac{\sigma}{c}\right)d\sigma$$

and conclude that, for this to hold for all A, it must be true that

$$\pi(\sigma) = \frac{1}{c}\pi\left(\frac{\sigma}{c}\right) \quad ;\text{for all } \sigma. \text{ Choosing } \sigma = c,$$

It follows that

$$\pi(c) = \frac{\pi(1)}{c}$$

Setting $\pi(1) = 1$; for convenience and noting that above equality must hold for all $c > 0$; it follows that a reasonable non-informative prior for a scale parameter is $\pi(\sigma) = \frac{1}{\sigma}$.

Observe that this is also an improper prior since $\int_0^\infty \frac{1}{\sigma}\,d\sigma = \infty$

13/08/23

Predictive Distribution

e.g. Weather forecast

→ predict the disease based on ~~collecting~~ previous/base knowledge/information

ie.

⇒ To understand why a patient can develop cancer, doctor sees various factors such as genes of patient, food intake, environment patient lives in; family health history; physic of patient etc.

$$y = B_0 + B_1 x_1 + B_2 x_2$$

y ↑ distribution; x_1 ↑ food habit; x_2 ↑ family history.

$x_1, \dots x_n \sim f(x|\theta)$

If we know the values of θ, then can we generate new distribution like $g(x|\theta)$, the more good estimation for new distribution the better it will fit on distribution graph of original $f(x|\theta)$

In Classical probability, we get point value for predictive probability so prediction becomes hard while in baysian statistics we get predictive distribution instead of point values of prediction.

Predictive Distribution - ~~If Sup~~

Suppose we are given random variable $x_1, x_2 x_3 \dots x_n$ from $f(x|\theta)$ and suppose $z_1, z_2, z_3, \dots z_n$ be predictive observation from the same distribution $f(x|\theta)$, then the previous sample is future or predictive

We have to find out $p(z|x)$ means we are given ~~pri~~ previous ~~and~~ data and want to find future data.

$x_1 \dots x_n \sim f(x|\theta)$

$z_1 \dots z_n \sim f(z|\theta)$

p ⇒ predictive

h ⇒ posterior

① ~~$h(\theta|x) = \frac{f(x|\theta).g(\theta)}{\int f(x|\theta).g(\theta)\,d\theta}$~~ → Joint → Marginal

② $\therefore p(z|x) = \frac{l(z,x)}{f(x)}$

$p(z|x) = \frac{p(z,x|\theta)}{p(x)}$

③ $p(z|x) = \frac{\int_{R_\theta} p(z,x|\theta).d\theta}{\int f(x|\theta).g(\theta).d\theta}$

$p(z|x) = \frac{f(z|\theta)\ f(x|\theta).g(\theta)}{f(x|\theta)\ g(\theta).d\theta}$

$$h(\theta|x) = \frac{f(x|\theta)\cdot g(\theta)}{\int_{R_\theta} f(x|\theta)\cdot g(\theta)\,d\theta}$$ → joint density / → marginal density

{ p ⇒ predictive, h ⇒ posterior.

$\therefore p(z|x) = \frac{l(z,x)}{f_1(x)} = \frac{p(z,x|\theta)}{p(x)}$

{ $f_1(x)$ is marginal distribution.

$$p(z|x) = \frac{\int p(z,x,\theta)\,d\theta}{p(x)} = \frac{\int p(z,x|\theta)\,g(\theta)\cdot d\theta}{\int_{R_\theta} f(x|\theta)\cdot g(\theta)\cdot d\theta}$$

$$p(z|x) = \frac{\int f(z|\theta)\, f(x|\theta)\cdot g(\theta)\cdot d\theta}{\int_{R_\theta} f(x|\theta)\cdot g(\theta)\cdot d\theta}$$ ← Posterior distribution of θ

$$p(z|x) = \int f(z|\theta)\cdot h(\theta|x)\cdot d\theta$$

$$p(z|x) = E_{\theta|x}\left(f(z|\theta)\right)$$

$\because \int \theta\cdot f(\theta|x)\,d\theta = E_{\theta|x}(\theta)$

θ is unknown and the problem is we don't even have sample to a infer about θ.

Here, $p(z|\theta)$ is the predictive distribution which gives the information about future data in the light of given data.

Q.] $x_1 \dots x_n$ are i.i.d's $\sim B(\theta)$ Bernoulli ; $\theta \sim Beta(\alpha,\lambda)$, prior also given posterior distribution $\theta|x \sim Beta(\sum x_i+\alpha, n-\sum x_i+\lambda)$. Find Prediction Distribution?

$x_1, x_2 \dots x_n \sim B(\theta)$

$B(\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ — ①

$\theta \sim Beta(\alpha,\lambda)$

$B(\alpha,\lambda) = \frac{1}{B(\alpha,\lambda)}\cdot\theta^{\alpha-1}\cdot(1-\theta)^{\lambda-1}$

$\theta|x \sim Beta(\sum x_i+\alpha, n-\sum x_i+\lambda)$

Note

$X \sim Beta(a,b)$

$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$

$0<x<1$

$a,b>0$

Joint Distribution (likelihood)

$$p(\theta, x) = P(\theta) \cdot P(x|\theta)$$

$$p(\theta|x) = \frac{P(\theta) \cdot P(x|\theta)}{\text{likelihood}}$$

Prior

likelihood fn

$$p(\theta|x) = \frac{1}{B(\alpha,\lambda)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\lambda-1} \cdot \theta^{\Sigma x_i} \cdot (1-\theta)^{n-\Sigma x_i}$$

$$= \frac{1}{B(\alpha,\lambda)} \cdot \theta^{(\alpha+\Sigma x_i - 1)} \cdot (1-\theta)^{\lambda+n-\Sigma x_i - 1}$$

$$\approx \frac{1}{B(\alpha+\Sigma x_i, \lambda+n-\Sigma x_i)} \cdot \theta^{\alpha+\Sigma x_i - 1} \cdot (1-\theta)^{\lambda+n-\Sigma x_i - 1}$$

Posterior Distribution : $B(\alpha+\Sigma x_i, \alpha+n-\Sigma x_i)$

Now, Predictive

$$p(z|x) = \int_0^1 f(z|\theta) \cdot p(\theta|x)\, d\theta$$

$$= \int_0^1 \theta^z \cdot (1-\theta)^z \cdot \frac{1}{B(\alpha+\Sigma x_i, \lambda+n-\Sigma x_i)} \cdot \theta^{\alpha+\Sigma x_i - 1} \cdot (1-\theta)^{\lambda+n-\Sigma x_i - 1}$$

$$= \frac{1}{B(\alpha+\Sigma x_i, \lambda+n-\Sigma x_i)} \int_0^1 \underbrace{\theta^{\alpha+z+\Sigma x_i - 1} \cdot (1-\theta)^{\lambda+n+z-\Sigma x_i - 1}}_{\text{beta fn}} \cdot d\theta$$

$$p(z|x) = \frac{B(\Sigma x_i + z + \alpha,\ \lambda + n - z - \Sigma x_i - 1)}{B(\alpha+\Sigma x_i,\ n+\lambda-\Sigma x_i)}$$

predictive Distribution.

14/07/23

Bayesian Interval

Confidence Interval → here interval is moving but θ is fixed.

(explanation on next page)

interval is fixed but the parameter 'θ' ~~is not~~ can move

The estimation of $\hat{\theta}$ is $\hat{\theta} = E(x_1 \cdots x_n)$ is one which minimizes the posterior expected loss.

So we need to workout for the probability that θ lies in the interval $[\theta_1, \theta_2]$ where $\theta_1 < \theta_2$, this interval which based on the posterior distribution $\theta|x$ is called credible interval.

$$1-\alpha = P[\theta_1 < \theta < \theta_2]$$

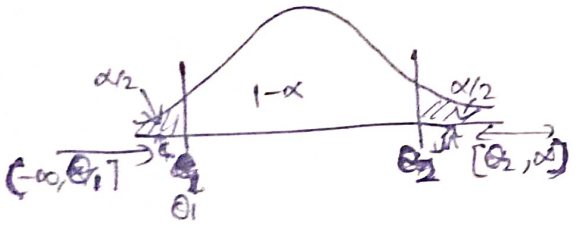
$$1-\alpha = \int_{\theta_1}^{\theta_2} p(\theta|x)\, d\theta \quad \text{— (1)}$$

we have infinite such solution which satisfy equation (1) so we have to make some strategy to choose the best amongst them.

(1) Equal Credible Interval

An equal tail $(1-\alpha)$ credible interval is given by

$$\frac{\alpha}{2} = \int_{-\infty}^{\theta_1} p(\theta|x)\, d\theta = \int_{\theta_2}^{\infty} p(\theta|x)\, d\theta \quad \rightarrow \text{from eqn (1)}$$



(2) Shortest Credible Interval

(reducing the credible interval)

⇒ To obtain the shortest $(1-\alpha)$ incredible interval, one has to minimize $I = \theta_2 - \theta_1$ which requires for $1-\alpha = [p(\theta_2|x) - p(\theta_1|x)]$ such that eqn (1) is satisfied ^ (it will be minimum when) $p(\theta_2|x) = p(\theta_1|x)$ — (2)

the ~~normal~~ interval θ_1, θ_2 will simultaneously satisfies equation (1) and (2) is called shortest $(1-\alpha)$ credible interval.

(3) Highest posterior density interval (HPD)

An interval $I = \theta_2 - \theta_1$ which satisfies following conditions simultaneously ⇒

① interval is shortest.

② $p(\theta|x)$ such that $\theta \in I > p(\theta|x)$ such that $\theta \notin I$

ie the posterior density inside at each point of the interval is greater than the posterior density at every point outside the interval, this ofcourse implies that the interval includes more probable values of θ and excludes the lesser ones.

Note: If posterior density is unimodal (not necessarily symmetric) the shortest credible interval and HPD interval are same

Credible Interval vs Confidence Interval ⇒

Credible Interval are a concept used in statistical inference and Baysian Statistics to estimate the uncertainty associated with a parameter or a set of parameter. Unlike frequentist estimation interval which provide a range of possible values for parameter based on repeated sampling credible interval provide a range of possible values based on available data & prior knowledge.

Credible Interval represents data and prior knowledge within which true values for the parameter is believed to lie with certain level of confidence based on data.

In frequentist terms, the parameter is fixed and the confidence interval is random. A credible interval is ~~random~~ simply an interval in the domain of the posterior distribution ~~or~~ within which an unobserved parameter value falls with. posterior possibility.

Q1] $g(\theta) = \frac{1}{\theta}$; $f(x,\theta) = \frac{1}{\theta} e^{-x/\theta}$; $x>0; \theta>0$

interval estimation

↓

All questions mostly in form of chi-square or z-test (distribution)

posterior $\propto$ likelihood $\times$ prior

$$\propto \frac{1}{\theta} \cdot \frac{1}{\theta^n} e^{-\Sigma x_i/\theta}$$

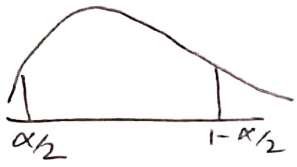
$$\propto \frac{1}{\theta^{1+n}} \cdot e^{-\Sigma x_i/\theta}$$

$$\frac{2\Sigma x_i}{\theta} \sim \chi^2_{2n}$$

$$P\left[\theta_1 < \frac{2\Sigma x_i}{\theta} < \theta_2\right] = 1-\alpha$$

$$P\left[\chi^2_{\alpha/2,2n} < \frac{2\Sigma x_i}{\theta} < \chi^2_{1-\alpha/2,2n}\right] = 1-\alpha$$

$$P\left[\frac{2\Sigma x_i}{\chi^2_{1-\alpha/2,2n}} < \theta < \frac{2\Sigma x_i}{\chi^2_{\alpha/2,2n}}\right] = (1-\alpha)$$



we can say that 100 $(1-\alpha)$% credible interval is

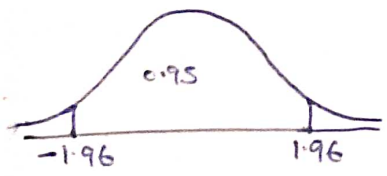
$$\left(\frac{2\Sigma x_i}{\chi^2_{1-\alpha/2,2n}}, \frac{2\Sigma x_i}{\chi^2_{\alpha/2,2n}}\right)$$

∴ therefore we can say random parameter θ lies 95% ~~th~~ time

b/w the fixed interval $\left(\frac{2\Sigma x_i}{\chi^2_{1-\alpha/2,2n}}, \frac{2\Sigma x_i}{\chi^2_{\alpha/2,2n}}\right)$ ← ~~both~~ interval is fixed & parameter is variable.

∴ This shortest credible interval.

Credible Interval$_{0.95}$ ⇒ ~~[0.95,~~ ~~[~~ [-1.96, 1.96]



~~Con~~ Credible Interval ⇒ In frequentist statistic out of 100 ~~the~~ time 95 time the unknown fixed parameter is supposed to lie b/w [-1.96, 1.96] ie. interval is random.

In Baysian Statistic, here parameter (θ) is random and interval is fixed. θ doesn't go out of interval [-1.96, 1.96] 95% ~~Out~~ of 100 times.

27/07/23

Critical situation
decide α (alpha) value

α = P(Type 1 error) = P(Reject H_0 | H_0 is true)

P(Type 2 error) = P(Accept H_1 | H_1 is true or H_0 is false)

** Baye's Testing →

Baye's Testing is quite straighforward as compared to classical definition here we accept null hypothesis if $p(H_0|x)$ is maximum and we accept alternate hypothesis if $p(H_1|x)$ is maximum

prior belief changes after observation of data

Posterior probability density of H_0 & H_1

$$p_0 = p(\theta \in H_0 | x) = p(H_0|x) = \int_{R_{H_0}} p(\theta|x)dx$$

$$p_1 = p(\theta \in H_1 | x) = p(H_1|x) = \int_{R_{H_1}} p(\theta|x)dx$$

$p_0 + p_1 = 1$ and $\pi_0 + \pi_1 = 1$

Prior Probability of H_0 & H_1

$\pi_0 = P(\theta \in H_0)$

$\pi_1 = P(\theta \in H_1)$

* Prior odds of H_0 against H_1 = $\frac{\pi_0}{\pi_1}$] when we convert this into probability $\frac{\pi_0}{\pi_1 + \pi_0}$

* Posterior odds of H_0 against H_1 = $\frac{p_0}{p_1}$

① If $\frac{\pi_0}{\pi_1} = 1$; then two hypothesis are same apriori

② If $\frac{\pi_0}{\pi_1} > 1$; apriori H_0 is preferred.

③ If $\frac{\pi_0}{\pi_1} < 1$; apriori H_1 is preferred.

Similarly,

① if $\frac{p_0}{p_1} = 1$; then two hypothesis are same a posteriori.

② if $\frac{p_0}{p_1} > 1$; a posteriori H_0 is preferred.

③ if $\frac{p_0}{p_1} < 1$; H_1 is preferred a posteriori

When data completely nullifies or deviate the prior, the posterior lies b/w them

* prior

* data

* posterior.

* Baye's Factor

we finally define baye's factor

$$B = \frac{p_0/p_1}{\pi_0/\pi_1}$$

this Baye's factor is the odds of H_0 against H_1

from other situation we can similarly define Baye's factor.

① $B = 1$; then both H_0 and H_1 are equally preffered

② $B < 1$; then Reject H_0 and accept H_1

③ $B > 1$; then Reject H_1 and Accept H_0

28/07/23

$$B = \frac{p_0/p_1}{\pi_0/\pi_1}$$

$$\frac{\pi_0}{\pi_1} B = \frac{p_0}{p_1}$$

$$\frac{\pi_0 B}{1-\pi_0} = \frac{p_0}{1-p_1} \quad [\because p_0 + p_1 = 1, \ \pi_0 + \pi_1 = 1]$$

$$p_0 = \frac{1}{1 + \left(\frac{1-\pi_0}{\pi_1}\right) B^{-1}}$$

$$p_1 = \frac{1}{1 + \left(\frac{1-\pi_1}{\pi_1}\right) B}$$

↑ check if B comes or B^{-1}

① Simple V/s Simple Hypothesis

$H_0 : \theta = \theta_0 \in (H)_0$

$H_1 : \theta = \theta_1 \in (H)_1$

$f(x|\theta_0)$

$f(x|\theta_1)$

$\pi_0 = g(\theta_0)$

$\pi_1 = g(\theta_1)$

$p_0 \propto f(x|\theta_0) g(\theta_0) \propto f(x|\theta_0)\pi_0$

$p_1 \propto f(x|\theta_1) g(\theta_1) \propto f(x|\theta_1)\pi_1$

B is for odds of $(H)_0$ against $(H)_1$

$$B = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{f(x|\theta_0)\cdot g(\theta_0) / f(x|\theta_1) g(\theta_1)}{g(\theta_0)/g(\theta_1)}$$

$$B = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

Conclusion ⇒ for simple vs simple hypothesis, Baye's Test is quite similar to likelihood ratio test in classical Inference.

---

* Simple vs Complex hypothesis

↓ no single value for hypothesis

$H_0 : \theta > 5$ ← composite, θ can take many values.

$H_1 : \theta = 5$ ↗ simple

* Simple vs Simple hypothesis

↓ Clearly defined hypothesis no double.

H_0: simple : $\theta = 5$

H_1 : simple : $\theta = 6$

(9.] $p(\lambda)$

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad ; \quad \lambda > 0, \; x = 0, 1, 2 \ldots \; \Big| \; \lambda = \lambda_0, \; \lambda = \lambda_1$$

$\lambda \sim \text{Gamma}(\alpha, \beta)$

$$g(\lambda) = \frac{\beta^\alpha}{\Gamma\alpha} e^{-\beta\lambda} \lambda^{\alpha-1} \quad ; \lambda > 0, \alpha, \beta > 0$$

$H_0 : \lambda = \lambda_0$

$H_1 : \lambda = \lambda_1$

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n}(x_i!)}$$

$p(\lambda|x)$

10/08/23 Reema Maam

Jeffery's ~~prior~~ Invariant Prior's

also known Non-informative prior or Vague or Ignorance prior

Proper in Bayesian means we are getting probability density fn

Improper prior multiplied by some constant gives proper prior

Rule ①

~~If~~ ~~θ~~ ~~(H)~~ $\in (-\infty, \infty)$

⇒ Jeffery suggested the following rules for choosing non-informative prior $\pi(\theta)$ for θ

proper prior

$\frac{1}{2\pi} \int_{-\infty}^{\infty} f(x|\theta)\, d\theta = 2\pi$

constant $\frac{1}{2\pi}$

Rule ① If parameter space $(H) = (-\infty, \infty)$

Assume θ is uniformly distributed

$\pi(\theta) =$ constant

or

$\pi(\theta) \propto 1$

This rule is invariant under any linear transformation.

~~$H:$~~ $U = a\theta + b$

Rule ② If $(H) = (0, \infty)$

Assume $\log\theta$ is uniformly distributed ie

$$\pi(\theta) \propto \frac{1}{\theta}$$

This rule is invariant under any exponential transformation.

$$U = \theta^c \; ; \quad c > 0$$

Rule 3 If Rule ① and Rule ② are members of a large family of prior then

$\pi(\theta) \propto |I(\theta)|^{1/2}$ ⟶ Jeffery's a general rule, no transformation restriction

where θ may be real or vector valued parameter and ~~$I(\theta)$~~

$$I(\theta) = -\cancel{E\left[\frac{\partial^2 \log(f(x|\theta))}{\partial\theta_i \partial\theta_j}\right]} - E\left[\frac{\partial^2 \log(f(x|\theta))}{\partial\theta_i \partial\theta_j}\right]$$ ⇒ Fisher's Information Matrix.

Q.1 If $X \sim N(\mu, \sigma^2)$; σ is known

$\pi(\mu) = c$

$\pi(\mu) \propto 1$

(2) If μ is known

$\pi(\sigma) \propto \frac{1}{\sigma}$ ⇒ using Rule 2 ; since $(0 < \sigma < \infty)$

(3) μ & σ are unknown

$\pi(\mu, \sigma^2)$

assuming μ, σ are independent

$\pi(\mu, \sigma) = \pi(\mu) \times \pi(\sigma|\mu)$

$= c \cdot \frac{1}{\sigma}$

$= \frac{c}{\sigma}$

$\therefore \pi(\mu, \sigma) \propto \frac{1}{\sigma}$

If μ & σ are independent.

then

$\pi(\mu, \sigma) = \pi(\mu) \cdot \pi(\sigma)$

$=$

(4) $X \sim \exp(\theta)$

$f(x|\theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$; $x, \theta > 0$

↑ scale parameter.

$\pi(\theta) \propto \frac{1}{\theta}$

(1) $x \sim N(\mu, \sigma^2)$, $\mu(-\infty, \infty)$; $\sigma(0, \infty)$

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

$$\log f(x|\theta) \propto = \propto c + \left(-\log \sigma\right)\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \Bigg| = c - \log\sigma.$$

$$\frac{\partial \log f(x|\theta)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

(1) $$\boxed{\frac{\partial^2}{\partial \mu^2} f(x|\theta) = -\frac{1}{\sigma^2}}$$

$\pi(\mu)$ = constant since σ is known.

(2) σ is unknown, μ is known.

$$\frac{\partial \log f(x|\theta)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$$

$$\frac{\partial^2 \log f(x|\theta)}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}$$

$$I(\sigma) = -E\left[\frac{\partial^2}{\partial \sigma^2} \log\left(f(x|\theta)\right)\right] = -E\left[\frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}\right]$$

$$= -\frac{1}{\sigma^2} + \frac{3}{\sigma^4} E(x-\mu)^2$$

$$= -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}$$

$$E(x-\mu)^2 = E(x-E(x))^2$$
$$= \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

③ when μ, σ are both unknown

$$= -E\begin{bmatrix} \frac{\partial^2 \log f(x|\mu)}{\partial \mu^2} & \frac{\partial^2 \log f(x|}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \log f(x|}{\partial \mu \partial \sigma} & \frac{\partial^2}{\partial \sigma^2} \log f(x|\sigma) \end{bmatrix}$$

$$= E\begin{bmatrix} -\frac{1}{\sigma^2} & \frac{-2(x-\mu)}{\sigma^3} \\ \frac{-2(x-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^2} \end{bmatrix}$$

$$\left(\frac{-2E(x-\mu)}{\sigma^3} \quad -\frac{2}{\sigma^3}\underbrace{\left(E(x) - \frac{E(x)}{\mu}\right)}_{0} \right)$$

$$= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{-2}{\sigma^2} \end{bmatrix}$$

$$I(\theta) = -E\left(\frac{1}{\sigma^2} \cdot \frac{2}{\sigma^2}\right)^{1/2} \propto \frac{1}{\sigma^2}$$

$$\pi(\theta, \sigma) = \frac{1}{\sigma}$$

$$\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$$

The above joint prior is different from $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma}$ obtained by using Rule ① and Rule ②. Box Tiao 1973, remarked that the extra factor $\frac{1}{\sigma}$ arises due to ignoring the prior independence b/w μ and σ^2.