Summary

- Frequentist framework. We choose estimators based on their sample properties. For example unbiasedness.

- Likelihood framework. We base inference on observed data. We choose estimators for the parameters that make the data most likely.

- Bayesian framework. We combine information before the experiment and information after the experiment in the posterior distribution through Bayes theorem.

3 Likelihood and Sufficiency

Now we focus on the important basic statistical concepts of likelihood and sufficiency, which are central to statistical inference. It also highlights how these fundamental concepts are used in the basic Bayesian approach.

3.1 Likelihood

Example:

- Suppose that in a sampling inspection of a large batch of items, 20 are chosen at random and classified as good or bad.

- θ is the proportion of bad items in the batch.

- the probability distribution of the number x of defective items in the batch is binomial

$$\Pr(X = x) = \binom{20}{x} \theta^x (1-\theta)^{20-x}, \qquad x = 0, \ldots, 20$$

- If we observe $x = 3$ then the probability of this event is

$$\binom{20}{3}\theta^3(1-\theta)^{17}$$

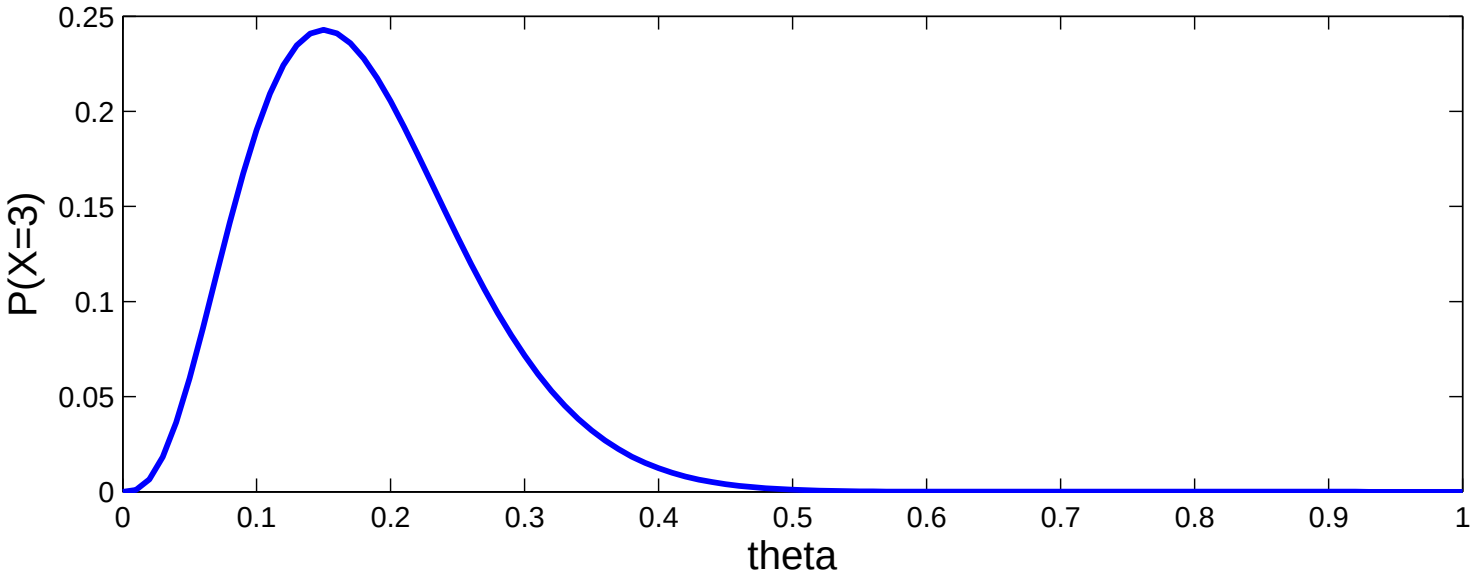
What does $x = 3$ tell us about θ?

If θ is unknown then it is helpful to tabulate the $\Pr(X = 3)$ for different value of θ

θ	Probaility	θ	Probability
0.03	0.018	0.19	0.218
0.05	0.060	0.21	0.192
0.09	0.167	0.25	0.134
0.13	0.234	0.29	0.082
0.15	0.243	0.31	0.062
0.17	0.236	0.35	0.032

Values of θ near to 0.15 are much more likely to have given rise to the observed result than, say, values of θ above 0.40

$$\Pr(X = 3) = \binom{20}{3}\theta^3(1-\theta)^{17}$$



We have plotted the probability of a particular result ($x = 2, n = 20$) for varying θ.

The function plotted is the Likelihood Function

Definition

The likelihood function is the joint probability of an observed sample, regarded as a function of the unknown parameters. The random variables (observations) are taken as fixed at their observed values.

If the observed data $x = (x_1, \ldots, x_n)$ from a statistical experiment have a joint probability density (or mass function, or a combination of the two, as appropriate) $p(x; \theta)$, then we regard this as a function of θ for fixed x.

We denote this function by $\mathcal{L}(\theta; x)$ or simply $\mathcal{L}(\theta)$.

It is defined over the parameter space Θ.

Simple Example: Suppose we have a random sample from and Exponential distribution of size 5: $X_i, i = 1, \ldots, 5$ i.i.d.

Before the experiment $\Longrightarrow$ Sampling distribution.

Vector of random variables

$$(X, X_2, X_3, X_4, X_5) \sim \lambda \exp\{\lambda x_1\} \lambda \exp\{\lambda x_2\} \lambda \exp\{\lambda x_3\} \lambda \exp\{\lambda x_4\} \lambda \exp\{\lambda x_5\}$$

$$= \lambda^5 \exp\{-\lambda \sum_{i=1}^{5} x_i\}$$

Suppose we observe data $(1.1, 2.3, 1.5, 3.4, 2.5) \Longrightarrow$ Likelihood function:

$$\mathcal{L}(\lambda) = \lambda^5 \exp\{-\lambda 10.8\}$$

Note that the only unknown is λ

In the special case (of most interest) when $X_1, \ldots, X_n$ is a random sample (i.e. the X_i are independent):

$$\mathcal{L}(\theta; x) = \begin{cases} \prod_{i=1}^{n} \Pr(X_i = x_i; \theta) & \text{if } X_i \text{ are discrete} \\ \prod_{i=1}^{n} p(x_i; \theta) & \text{if } X_i \text{ are continuous} \end{cases}$$

- A data set x contains information about the parameter θ only if its joint probability $p(x;\theta)$ depends on the parameter value.

- However, it is possible that even when this is true for the data set as a whole, particular data components may not contribute to that dependence.

- For example, in example 1.1 the data are from 10 independent replicates of an experiment. Those replicates where the coin results in tails give $R = 0$ and each contributes a factor of $\frac{1}{2}$ to the joint probability mass function. The independence of all the replicates means that there is no dependence of the other data values (whose probabilities *do* depend on θ) on those replicates for which $R = 0$. Thus, no information about θ is gained by observing $R = 0$.

- Suppose that the observed data x can be partitioned into two subsets x_1 and x_2, so that their joint probability can be written as the product $p(x_1;\theta)p(x_2 \mid x_1;\theta)$.

- If the marginal probability $p(x_1;\theta)$ of x_1 does not depend on θ then the values x_1 will still contribute to inference about θ via their role in $p(x_2 \mid x_1;\theta)$ unless X_1 and X_2 are independent ($p(x_2 \mid x_1;\theta) = p(x_2 \mid \theta)$).

- This latter is the case applying in example 1.1.

- The likelihood represents the probability of the data actually observed if θ is the true parameter value.

- it can be regarded as a measure of the compatibility/plausibility of parameter values with the observed data

- if θ_1 has a greater likelihood than θ_2 then θ_1 is, in some sense, a more 'likely' value for θ than θ_2.

- Note however that more 'likely' does not mean more probable, unless we work in a Bayesian context in which θ is treated as a random variable.

- The interpretation of the likelihood function is to give some measure of the relative plausibility of different sets of unknown parameters giving rise to the observed data.

Maximum Likelihood Principle

Given the likelihood function $\mathcal{L}(\theta)$, choose as an estimate of θ that value $\widehat{\theta}$ which maximises the likelihood function:

$$\widehat{\theta} = \sup_{\theta \in \Theta} \mathcal{L}(\theta; x)$$

This is known as the maximum likelihood estimate (mle).

In practice, because the logarithm is an increasing function of its argument, it is equivalent and often more convenient to choose θ to maximise the log-likelihood function, $\ell(\theta) = \ln \mathcal{L}(\theta)$.

Example 3.1

Suppose we observe a sequence of n independent Bernoulli trials with (unknown) success parameter, θ, in which there are r successes. The likelihood function is given by

$$\begin{aligned} \mathcal{L}(\theta) &= \theta^r (1-\theta)^{n-r}, \qquad \text{for } 0 \leq \theta \leq 1 \\ \ell(\theta) &= r \ln \theta + (n-r) \ln(1-\theta) \end{aligned}$$

$\Longrightarrow$ the mle is simply the observed proportion r/n. Values of θ around r/n are the 'most likely', but we should not interpret these values as the 'most probable'.

- Note that the likelihood function is the same if the data were obtained by negative binomial sampling, i.e. if we had sampled trials until the rth success were obtained and had needed n trials.

- the sampling method has no bearing on the value of the mle and it is only the number of trials and the number of successes that matter.

- This is because the likelihood approach focuses on the probability (or probability density) of the data actually observed and does not take the rest of the sample space into account.

It is important to remember in maximising the likelihood or log-likelihood that differentiation gives only local maxima and minima, so that even if the second derivative $d^2\ell(\theta)/d\theta^2$ is negative (corresponding to a local maximum, rather than minimum) it is still possible that the global maximum is elsewhere. The global maximum will be either a local maximum or achieved at the boundary of the parameter space.

An important result is that if we reparameterise the distribution by using particular functions of the original parameters, then the maximum likelihood estimates of the new parameters are the corresponding functions of the maximum likelihood estimates of the original parameters:

$$\widehat{\theta} \text{ is M.L.E. of } \theta \Rightarrow g(\widehat{\theta}) \text{ M.L.E. of } g(\theta)$$

Regarded as random variables, maximum likelihood estimators can be shown to have good properties; these properties will be discussed later.

Note: if two different sets of observations determine the same likelihood, then the same value of the M.L.E. of θ will be obtained for both sets.

Suppose two different samples $x = (x_1, \ldots, x_n)$ and $\mathrm{y} = (y_1, \ldots, y_n)$ determine likelihood function which are proportional to each other:

$$\mathcal{L}(\theta; x) = c\mathcal{L}(\theta; \mathrm{y})$$

Then the MLE for θ will be the same regardless of whether x or y is considered.

This is known as the Likelihood Principle.

Likelihood Principle

The information brought by observations x about θ is entirely contained in the likelihood function $\mathcal{L}(\theta; x)$. Moreover , if x and y are two sets of observations depending on the same parameter θ, such that there exists a constant c satisfying

$$\mathcal{L}(\theta; x) = c\mathcal{L}(\theta; \mathrm{y})$$

for every θ, they bring the same information about θ and must lead to identical inference on θ.

Example

A statistician must estimate the prportion θ of defective items in a large manufactured lot. The statistician is informed that 10 items were randomly selected form the lot and 2 were found defective. If no additional information is provided on the experiment, two probability models at least can be proposed.

Possible experiments

1. Binomial Sampling: a sample of 10 items has been selected at random and 2 items were found defective. ($X \sim$ Bin$(10, \theta)$)

2. Negative-Binomial Sampling: Items had been selected at random from the lot until 2 defective item were found and it was found that a total number of 10 items had been selected. ($X \sim$ Neg-Bin$(2, 1-\theta)$)

In other words the random quantity in the experiment can be either 2 or 10.

Important: For both the experiments the likelihood is proportional to $\theta^2(1-\theta)^8, 0 \leq \theta \leq 1$, and inference on θ should be identical for both models.

3.2 Sufficiency

In Example 3.1, suppose that the data available for estimating θ were just the total number, R, of successes obtained and the number n of Bernoulli trials, where n is assumed known. Since the Bernoulli trials were independent, it is clear that there was no benefit in knowing the sequence of results of the individual trials. In other words, given independence, the order of these results provides no information about θ.

More formally, we can see that the conditional distribution of the individual results, $X_1, \ldots, X_n$ say, given R, does not depend on θ.

For $r = 0, \ldots, n$,

$$\begin{aligned} \mathsf{P}(X_1 = x_1, \ldots, X_n = x_n \mid R = r) &= \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{r}\theta^r(1-\theta)^{n-r}} \\ &= \frac{\theta^r(1-\theta)^{n-r}}{\binom{n}{r}\theta^r(1-\theta)^{n-r}} = \frac{1}{\binom{n}{r}}, \end{aligned}$$

if $x_1 + \cdots + x_n = r$, and $\mathsf{P}(X_1 = x_1, \ldots, X_n = x_n \mid R = r) = 0$ otherwise.

$\Longrightarrow$ only the number of successes contains information on the θ.

Recall

Given a sample X, A statistic $T(X)$ is an observable function of the observation vector X, i.e. $T(X)$ is directly computable from the data vector X. A statistic can be univariate or can be a vector.

Example of common statistics: Sample mean and sample variance, sample moments

In general, certain aspects of the data are relevant to making decisions (such as estimation and hypothesis testing) about a parameter of interest and others are not.

If a sample X has a distribution that depends on θ, then a statistic $T(X)$ (which may be vector-valued) is sufficient for θ if the conditional distribution of X given $T(X)$ is independent of θ.

In this case we feel confident that $T(X)$ contains all the information about θ that is contained in the sample.

Definition: A statistic $T(X)$ is sufficient for a parameter, θ, if and only if the conditional distribution, $p(x \mid t; \theta)$, of X given $T(x) = t$, does not depend on θ.

Since we can factorise the joint distribution of the data in the form

$$p(x; \theta) = p(x, t; \theta) = p(x \mid t; \theta)\, p(t; \theta)$$

it is clear that if the first factor does not depend on θ, then the sufficient statistic, T, contains all the information about θ in the data X.

The principle of sufficiency says that inferences about θ should involve the data only through the sufficient statistic T.

Sufficiency Principle

If two sets of observations x and y lead to the same value of a sufficient statistic T, i.e. $T(x) = T(\mathrm{y})$, then they must lead to the same inference on θ.

Notation: We write the statistic, which is a function of the random variable X and thus random, as $T(X)$. Its value is thus $T(x) = t$. However, it is often simpler (as above) to abbreviate the random statistic to T.

It follows immediately from the definition that if T is sufficient for θ then

(a) so is any invertible function of T, and

(b) so is (T, U) where $U(X)$ is any statistic, so a sufficient statistic is not unique.

In fact the full sample X is itself a sufficient statistic (so a sufficient statistic always exists) but provides no reduction of the data.

By concentrating on the data only through a suitably chosen sufficient statistic, the dimensionality of the problem can often be significantly reduced (to that of T rather than that of X).

A sufficient statistic that is a function of every other sufficient statistic is called a minimal sufficient statistic, and provides the maximum reduction of the data.

A minimal sufficient statistic most efficiently captures all possible information about the parameter θ.

Note that while both θ and T are, in general, vectors, their dimensions need not be the same and, in particular, if θ is a single (scalar) value, there may not exist any scalar sufficient statistic (examples will be given later).

A single sufficient statistic is usually, but not necessarily, minimal sufficient.

In Example 3.1 the maximum likelihood estimator R/n is a sufficient statistic for the probability θ of success.

Since

$$p(x;\theta) = p(x \mid t)p(t;\theta) = \mathcal{L}(\theta)$$

$\Longrightarrow$ maximising the likelihood is equivalent to maximising $p(t;\theta)$ with respect to θ

Hence a maximum likelihood estimator is necessarily a function of the (minimal) sufficient statistic.

Example 3.2

X, Y are iid $\mathsf{N}(\mu, \sigma^2)$ with σ known.

Define $T = X + Y$: is T sufficient for μ?

Determine $p_{X,Y|T}(x, y \mid t, \mu)$: does this distribution depends on μ?

- $p_{X,Y}(x, y; \mu)$ is equal to the product of 2 Normal distributions

$$p_{X,Y}(x, y; \mu) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[(x-\mu)^2 + (y-\mu)^2\right]\right\}$$

- T is a linear combination of Normals so $T \sim \mathsf{N}(2\mu, 2\sigma^2)$

$$p_T(t; \mu) = \frac{1}{\sqrt{2\pi 2\sigma^2}} \exp\left\{-\frac{1}{4\sigma^2}\left[(t-2\mu)^2\right]\right\}$$

The conditional density of X, Y given $x + y = t$ is

$$p_{X,Y|T}(x, y \mid x + y = t) = \frac{\frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[(x-\mu)^2 + (t-x-\mu)^2\right]\right\}}{\frac{1}{\sqrt{2\pi 2\sigma^2}} \exp\left\{-\frac{1}{4\sigma^2}\left[(t-2\mu)^2\right]\right\}}$$

$$= \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{1}{4\sigma^2}\left[2(x-\mu)^2 + 2(t-x-\mu)^2 - (t-2\mu)^2\right]\right\}$$

$$= \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{1}{4\sigma^2}\left[4x^2 + t^2 - 4tx\right]\right\}$$

which does not depend on $\mu \Longrightarrow T$ is sufficient for μ.

End of Example 3.2 ■

- In the above example, we started with a statistic T and then derived the conditional distribution of X given T to show that it is a sufficient statistic.

- If the conditional distribution had turned out to depend on the unknown parameter then we would have had to guess another form for a sufficient statistic and start again.

- In general, finding sufficient statistics could be difficult

- Fortunately, however, there is an 'automatic' way of finding sufficient statistics: Neyman's Factorisation Theorem

Neyman's Factorisation Theorem

A statistic T is sufficient for θ if and only if there is a factorisation of the joint distribution of the data of the form

$$p(x;\theta) = h(x)g(T(x);\theta)$$

where $g(T(x);\theta)$ involves the data x only through $T(x)$ and $h(x)$ does not depend on θ.

Proof: (*discrete case only*, **ONLY FOR INTERESTED STUDENTS**)

Assume T is sufficient. Then, by definition

$p(x \mid T(x))$ does not depend on θ and is a function only of x

Then

$$p(x;\theta) = p(x \mid T(x) = t)p(t;\theta) = h(x)g(T(x);\theta)$$

Assume that $p(x;\theta) = h(x)g(T(x);\theta)$ holds. Then,

$$\Pr(X = x \mid T(X) = t) = \frac{\Pr(X = x; T(x) = t)}{\Pr(T(X) = t)}$$

$$\Pr(T(X) = t) = \sum_{\{x:T(x)=t\}} \Pr(X = x;\theta) = \sum_{\{x:T(x)=t\}} h(x)g(T(x);\theta)$$

$$= g(t;\theta) \sum_{\{x:T(x)=t\}} h(x)$$

Note that $\Pr(X = x \mid T(x) = t) = 0$ if $T(x) \neq t$.

Since T is a function of X, $\Pr(X = x; T(x) = t) = \Pr(X = x;\theta)$.

Hence:

$$\begin{aligned}
\Pr(\boldsymbol{X} = x \mid \boldsymbol{T}(\boldsymbol{X}) = t) &= \frac{\Pr(\boldsymbol{X} = x; \boldsymbol{\theta})}{g(t;\boldsymbol{\theta}) \sum_{\{x:\boldsymbol{T}(x)=t\}} h(x)} \\
&= \frac{h(x) g(t;\boldsymbol{\theta})}{g(t;\boldsymbol{\theta}) \sum_{\{x:\boldsymbol{T}(x)=t\}} h(x)} \\
&= \frac{h(x)}{\sum_{\{x:\boldsymbol{T}(x)=t\}} h(x)}
\end{aligned}$$

which does not depend on θ.

$\Longrightarrow$ by definition T is sufficient.

End of Proof ■

Example 3.3

$X_i \overset{iid}{\sim} \mathsf{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ with μ, σ^2 unknown

$$
\begin{aligned}
p(x; \mu, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu)^2\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\overline{x} - \mu)^2}{2\sigma^2}\right\}
\end{aligned}
$$

where $\overline{x} = \sum_{i=1}^{n} x_i/n$ and $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$.

Take $T = (\overline{X}, S^2)$ and apply the factorisation the theorem with $h(x) = 1 \Longrightarrow T$ is sufficient for μ, σ^2

- If σ^2 is known and μ unknown $\Longrightarrow$ $\overline{X}$ is sufficient for μ. Set

$$h(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}$$

$$g(t;\mu) = \exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\}$$

- if μ is known and σ^2 is unknown, the inknown σ^2 enters both terms in the exponential. $(\overline{X}, S^2)$ are jointly sufficient for σ^2. However $\sum_{i=1}^n (X_i - \mu)^2$ is a single sufficient statistic for σ^2

End of Example 3.3 ■

Comments

1. We see from this example that in the case of a normal population, no information is lost by retaining only the sample mean and variance. This is not true in general.

2. Example 3.3 illustrates the fact that if $T = (T_1, \ldots, T_l)$ is sufficient for $\theta = (\theta_1, \ldots, \theta_k)$ it need not follow that each T_r is sufficient for a single θ_j. The converse, that if each T_r is sufficient for a single θ_j then $T = (T_1, \ldots, T_l)$ is sufficient for $\theta = (\theta_1, \ldots, \theta_k)$, is always true.

3. Normally the dimension of a sufficient statistic T is the same or greater than that of θ, but examples can be constructed where it is less.

Example 3.4

$X_i \overset{iid}{\sim}$ Uniform$[0, \theta]$, $i = 1, \ldots, n$ and θ unknown.

$$p(x;\theta) = \begin{cases} \frac{1}{\theta^n} & 0 \leq x_i \leq \theta \text{ i.e. } \max x_i \leq \theta \text{ and } \min x_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Define $x_{(1)} = \min x_i$, $x_{(n)} = \max x_i$ (order statistics) and the Heaviside function

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Then

$$p(x;\theta) = \frac{1}{\theta^n} H(x_{(1)}) H(\theta - x_{(n)})$$

$\Longrightarrow$ $x_{(n)}$ is sufficient by factorisation theorem.

End of Example 3.4 ■

Example 3.5

Let $Y \sim \text{Exp}(\lambda)$ and $X = Y + \mu$ (shifted exponential). Then

$$p(x;\lambda,\mu) = \begin{cases} \lambda \exp\left(-\lambda(x-\mu)\right) & x \geq \mu \\ 0 & \text{otherwise} \end{cases}$$

Assume $X_1, \ldots, X_n$ i.i.d. as X:

$$\begin{aligned} p(x_1,\ldots,x_n;\lambda,\mu) &= \lambda^n \exp\left\{-\lambda \sum x_i + n\mu\lambda\right\}, \qquad \mu \leq x_i, \forall i \\ &= \lambda^n \exp\left\{-\lambda \sum x_i + n\mu\lambda\right\} H(x_{(1)} - \mu) \end{aligned}$$

where $X_{(1)} = \min X_i$ and $H(x)$ is the Heaviside function.

From factorisation theorem:

- $(\sum X_i, X_{(1)})$ are jointly sufficient for (μ, λ)
- if λ is known, $X_{(1)}$ is sufficient for μ
- if μ is known, $\sum X_i$ is sufficient for λ

End of Example 3.5 ■

3.3 Bayesian Methods

The Bayesian approach involves

- the specification of a prior distribution, $\pi(\theta)$, for θ, which represents our 'degree of belief' about the state of nature θ before the experiment has been performed and the data observed.

- since θ is regarded as a random variable, the distribution of the data X, $p(x;\theta)$, is the conditional distribution of X given θ: $p(x \mid \theta)$

- After the data have been observed, the prior distribution $\pi(\theta)$ of θ is updated to a posterior distribution $\pi(\theta \mid x)$ using Bayes' theorem:

$$\pi(\theta \mid x) \propto \pi(\theta) p(x \mid \theta)$$

Inference about θ is based on this posterior distribution, where interest focuses on this as a function of θ.

- If the prior distribution is a constant (a uniform prior) then the posterior distribution has exactly the same shape as the likelihood function:

$$\pi(\theta \mid x) \propto \text{ const. } p(x \mid \theta)$$

- It is a probability distribution and can be used to make probability statements about θ.

Discrete and Continuous Prior Distributions

- In some problem the parameter θ can take only only a finite number of values or, at most, an infinite sequence of different values. The prior distribution $\pi(\theta)$ will there for be a discrete distribution.

- In other problems, the parameter θ can take any value on the real line or in some interval of the real line, then $\pi(\theta)$ is a continuous prior distribution.

Example: Fair or Two Headed Coin

θ: probability of obtaining head when a certain coin is tossed.

Suppose we know that the coin either is fair ($\theta = 0.5$) or has head on each side ($\theta = 1$).

If the prior probability that the coin is fair is p (e.g. $p = 0.3$), then

$$\begin{aligned} \pi(0.5) &= p \\ \pi(1) &= 1 - p \end{aligned}$$

Question: What happens to the posterior distribution of θ id you observe a tail in 5 flips of the coin? What if you do not observe any tail?

Example 3.6

Suppose we observe r successes in n Bernoulli trials, with probability θ of success.

Assume that our prior knowledge about θ can be represented by a Beta(4,6) distribution.

Prior distribution:

$$\pi(\theta) = \frac{\theta^3(1-\theta)^5}{\mathbf{B}(4,6)}$$

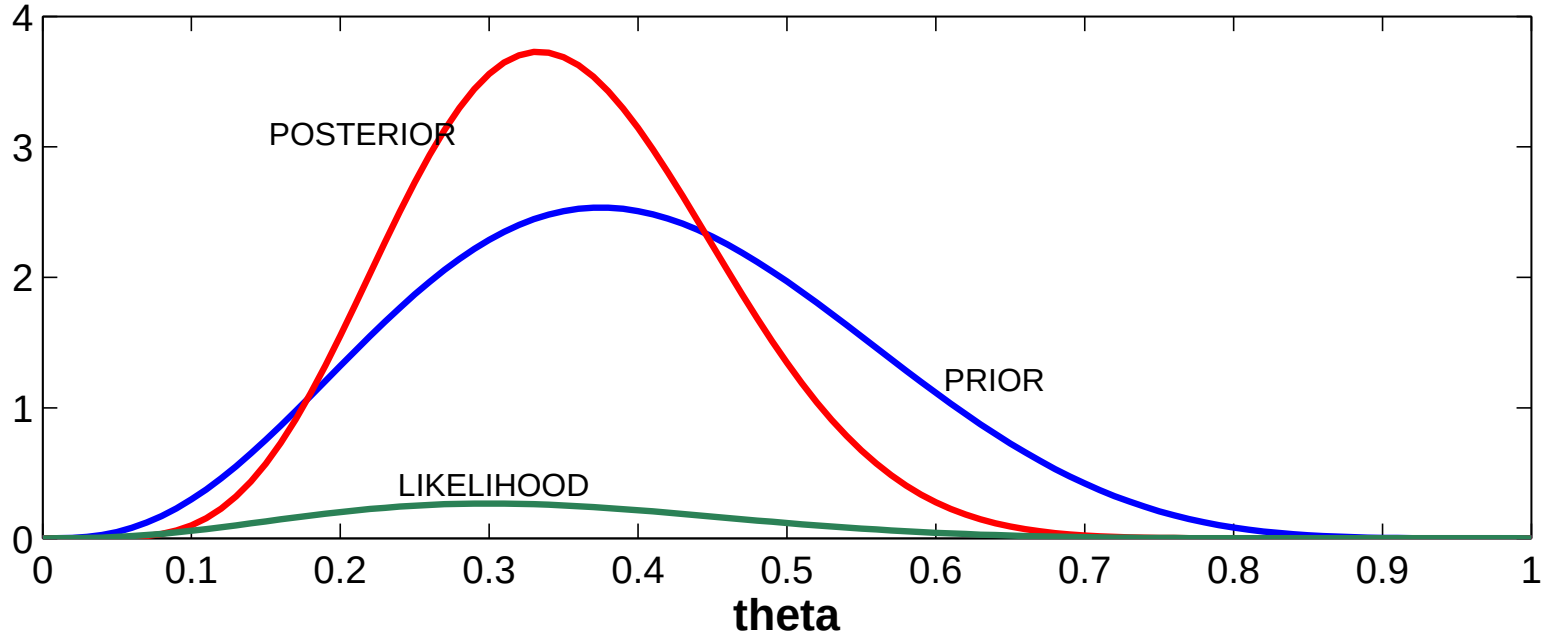
Likelihood:

$$p(r \mid \theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$$

Posterior distribution:

$$\pi(\theta \mid r) \propto \theta^3(1-\theta)^5\theta^r(1-\theta)^{n-r} = \mathbf{Beta}(4+r, 6+n-r)$$

If $n = 10$ and $r = 3$, then the posterior distribution is a Beta(7,13)



The Beta(α, β) has mean $\alpha/(\alpha+\beta)$ so a priori θ has mean 0.4. After observing the data, the posterior mean is 0.35.

Note that prior and posterior distribution are from the same parametric family.

End of Example 3.6 ■

- If more data became available then this posterior distribution can be further updated, taking the role of the 'prior' distribution for θ before the new data are incorporated, to give a revised posterior distribution after both sets of data have been taken into account.

- The final posterior distribution is the same, regardless of whether data x_1 are incorporated first and then x_2 or *vice versa*.

- We start with a prior $\pi(\theta)$ for θ. We observe x_1 and obtain posterior $p(\theta \mid x_1) \propto \pi(\theta)p(x_1 \mid \theta)$

- The posterior $p(\theta \mid x_1)$ serves as the prior for θ when the second experiment is to be performed. After we observe x_2, the posterior for θ is

$$\begin{aligned} \pi(\theta \mid x_1, x_2) &\propto p(x_2 \mid \theta)p(\theta \mid x_1) \\ &\propto p(x_2 \mid \theta)\pi(\theta)p(x_1 \mid \theta) \end{aligned}$$

- we can continue in this way updating the posterior distribution as more samples are collected.

Sufficiency in a Bayesian context

If $T(X)$ is sufficient for θ then the probability $p(x \mid \theta)$ factorises so that

$$p(x \mid \theta) = p(x \mid t)\, p(t \mid \theta)$$

$\Longrightarrow$ the posterior distribution satisfies

$$\pi(\theta \mid x) \propto p(x \mid \theta)\, \pi(\theta) = p(x \mid t)\, p(t \mid \theta)\, \pi(\theta) = p(x \mid t)\, \pi(\theta \mid t)\, p(t) \propto \pi(\theta \mid t)$$

The statistic $T(X)$ is sufficient for a parameter, θ , if and only if $\pi(\theta \mid x) = \pi(\theta \mid t)$.

Knowledge of T gives the same inferences about θ as if we knew the full data x.