STAT0008 Lecture 2
Sufficiency and Decision Theory

Dr. Aidan O'Keeffe

Department of Statistical Science
University College London

8th October 2018

Outline

- Sufficiency
  - Sufficient statistics
  - Factorisation criterion
  - Minimal Sufficiency
- (Recap of) Bayesian methods
- Statistical Decision Theory
  - Loss functions
  - Frequentist, likelihood and Bayesian approaches

Definition: Sufficient Statistic

Suppose that $T = T(X_1, \ldots, X_n)$ is a statistic formed from the sample $X_1, \ldots, X_n$ where the probability distribution of $(X_1, \ldots, X_n)$ is parameterised by $\theta \in \Theta$. The statistic T is said to be **sufficient** for θ if the conditional probability distribution of $X_1, \ldots, X_n$ given $T = t$ does not depend on θ. That is

$$f(x_1, \ldots, x_n \mid T = t)$$

does not depend on θ where $f(x_1, \ldots, x_n \mid T = t)$ denotes the probability density (or mass) function of $X_1, \ldots, X_n$ conditional on $T = t$.

Sufficient Statistics

In other words, if $T(X)$ is sufficient for θ then knowing more about the sample of data than $T(X) = t$ will be of no use in making inference about θ.

The maximum likelihood estimator (where it exists) is always a function of a sufficient statistic.

We can find sufficient statistics by using the **factorisation criterion**.

Factorisation Criterion

Definition: Factorisation Criterion

The statistic T is sufficient for θ if and only if the joint density $f(\mathbf{x};\theta)$ (i.e. the **likelihood function**) can be expressed as

$$f(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x})$$

where $g(.)$ and $h(.)$ are some non-negative functions.

In words, the joint density (likelihood function) can be written as a product of two functions g and h, where g depends on θ and the sufficient statistic, but h does not depend on θ.

Factorisation Criterion - Proof

We'll prove the discrete case only. Let $f(\mathbf{x};\theta) = \mathbb{P}(\mathbf{X} = \mathbf{x};\theta)$.

$\Longrightarrow$

Suppose T is a sufficient statistic for θ and let $T(\mathbf{x}) = t$. Then

$$\mathbb{P}(\mathbf{X} = \mathbf{x};\theta) = \mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t, \theta)\mathbb{P}(T(\mathbf{X}) = t;\theta)$$

But since T is sufficient for θ, we can write

$$\mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t, \theta) = \mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t)$$

because sufficiency implies that the conditional distribution of $\mathbf{X}$ given T is independent of θ. Hence

$$\begin{aligned}\mathbb{P}(\mathbf{X} = \mathbf{x};\theta) &= \mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t)\mathbb{P}(T(\mathbf{X}) = t;\theta) \\ &= h(\mathbf{x})g(T(\mathbf{x}), \theta).\end{aligned}$$

Factorisation Criterion - Proof

$\Longleftarrow$

Now, suppose that $\mathbb{P}(\mathbf{X} = \mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. Then

$$\mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) = \frac{\mathbb{P}(\mathbf{X} = \mathbf{x} \cap T(\mathbf{X}) = t)}{\mathbb{P}(T(\mathbf{X}) = t)} = \frac{g(t, \theta)h(\mathbf{x})}{\mathbb{P}(T(\mathbf{X}) = t)}$$

We note that

$$\begin{aligned}
\mathbb{P}(T(\mathbf{X}) = t) &= \sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} \mathbb{P}(\mathbf{X} = \mathbf{x}; \theta) \\
&= \sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} g(T(\mathbf{x}), \theta)h(\mathbf{x}) \\
&= g(t, \theta) \sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} h(\mathbf{x})
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) &= \frac{g(t, \theta)h(\mathbf{x})}{g(t, \theta) \sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} h(\mathbf{x})} \\
&= \frac{h(\mathbf{x})}{\sum_{\{\mathbf{x}:T(\mathbf{x})=t\}} h(\mathbf{x})}
\end{aligned}$$

which does not depend on θ, so T is sufficient for θ.

Sufficiency: Example 1

Suppose that $X_1, \ldots, X_n$ are iid random variables where $X_i \sim \text{Poi}(\lambda)$ ($i \in \{1, \ldots, n\}$). Show that $\sum_{i=1}^{n} X_i$ is sufficient for λ.

Sufficiency: Example 2

Suppose that $X_1, \ldots, X_n$ are iid random variables where $X_i \sim \mathcal{N}(\mu, \sigma^2)$ $(i \in \{1, \ldots, n\})$. Find sufficient statistics for μ and σ^2.

Sufficiency: Example 2

Comments

This example illustrates the fact that if $\mathbf{T} = (T_1, \ldots, T_k)$ is sufficient for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, it need not follow that each T_j is sufficient for a single θ_j. Conversely, if each T_j is sufficient for a single θ_j then $\mathbf{T} = (T_1, \ldots, T_k)$ is guaranteed to be sufficient for $\boldsymbol{\theta}$.

In general, sufficient statistics are helpful for inference since they often reduce the dimensionality of a problem.

For example, where $X_1, \ldots X_n$ are iid Poi(λ), we can use the one-dimensional sufficient statistic $\sum_{i=1}^{n} X_i$ to make inference about λ, having started with the n-dimensional 'statistic' $(X_1, \ldots, X_n)$ - i.e. our data set!

Sufficiency Principle

The **sufficiency principle** states that if two sets of data $\mathbf{x}$ and $\mathbf{y}$ result in the same value of a sufficient statistic T (i.e. $T(\mathbf{x}) = T(\mathbf{y})$, where T is sufficient for θ) then these sets of data **must** lead to the same inference on θ.

Notes

If T is sufficient for θ then

(a) any invertible function of T is also sufficient for θ.

(b) (T,U), where $U(\mathbf{X})$ is any statistic constructed using $\mathbf{X}$, is also sufficient for θ.

So, a sufficient statistic is not unique! (e.g. Poisson example, both $\sum_{i=1}^{n} X_i$ and $\bar{X}$ are sufficient for λ.)

Minimal Sufficiency

A sufficient statistic that is a function of all other sufficient statistics, for a given parameter θ, is called a **minimal sufficient** statistic.

Formally, if $\mathbf{x}$ and $\mathbf{y}$ are two samples of data with joint density functions $f(\mathbf{x};\theta)$ and $f(\mathbf{y};\theta)$, respectively, then T is minimal sufficient for θ if and only if

$$\frac{f(\mathbf{x};\theta)}{f(\mathbf{y};\theta)} \text{ is independent of } \theta \iff T(\mathbf{x}) = T(\mathbf{y}).$$

Minimal Sufficiency - Example

Suppose that $X_1, \ldots, X_n$ are iid random variables such that $X_i \sim \text{Exp}(\theta)$. Find a minimal sufficient statistic for θ.

Minimal Sufficiency - Example

Bayesian Methods

Suppose that $X_1, \ldots, X_n$ are random variables with some probability distribution that has parameter(s) θ.

$$\text{i.e. each } \; X_i \sim \mathcal{D}(\theta).$$

Up to now, we have considered the parameter(s) θ to be fixed (but unknown) quantities and our goal has typically been to estimate θ. However, there are other assumptions that can be made about θ...

In a **Bayesian framework** we assume that distribution parameter(s), θ, are random variables. The inference process is a follows...

Bayesian Methods

Suppose, as before, that we consider a statistical model for data $X_1, \ldots, X_n$ where $X_i \sim \mathcal{D}(\theta)$. Let $f(\mathbf{x}; \theta)$ be the joint density of $X_1, \ldots, X_n$, conditional on θ. **Note**: We've done nothing special yet, we have simply assumed a probability distribution for our data.

1. A **prior probability distribution** (commonly called a 'prior' or 'prior distribution') is specified for the parameter(s) of interest θ. The prior distribution is usually denoted $\pi(\theta)$.

2. Now, suppose that a sample of data $\mathbf{x} = (x_1, \ldots, x_n)$ is observed. We construct a likelihood function using the specified model for the data and write this as

$$\mathcal{L}(\theta \mid \mathbf{x}) = f(\mathbf{x}; \theta).$$

3. We use **Bayes' theorem** to combine the prior distribution and likelihood function and form a **posterior distribution** for θ. Bayes' theorem implies that

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

We write this as

$$\pi(\theta \mid \mathbf{x}) \propto \pi(\theta) \times \mathcal{L}(\theta \mid \mathbf{x}).$$

We see that the prior distribution, $\pi(\theta)$, has been combined with the likelihood function (information about θ from the data sample) to produce the **posterior distribution** $\pi(\theta \mid \mathbf{x})$.

Bayesian Methods

The **prior** distribution, $\pi(\theta)$, represents our beliefs about θ *before* (i.e. *prior to*) the observation of the data $\mathbf{x}$.

The **posterior** distribution, $\pi(\theta \mid \mathbf{x})$, represents our updated beliefs about θ, having observed the set of data $\mathbf{x}$.

Inference about θ can be made by examining the posterior distribution and its properties.

Since the posterior is a probability distribution, we can use statements of probability to describe our beliefs (or uncertainties!) about θ. For example, $\mathbb{P}(\theta > 5 \mid \mathbf{x})$ or $\mathbb{E}(\theta \mid \mathbf{x})$ etc.

Bayesian Inference - Example

In a British political constituency, a pollster wants to know about the porportion of eligible voters who will vote *Labour* in an upcoming general election. The pollster plans to sample 100 people who live in the constituency and ask them how they plan to vote. We shall assume that there is no non-response.

We define

$$X_i = \text{ Voting intention of } i^{\text{th}} \text{ sampled voter, where}$$

$$X_i = \begin{cases} 0 & \text{if person } i \text{ will not vote } \textit{Labour}; \\ 1 & \text{if person } i \text{ will vote } \textit{Labour}. \end{cases}$$

Our assumed model for the data is

$$X_i \sim \text{Bernoulli}(\theta) \text{ independently for } i = 1, \ldots, 100.$$

where θ is the true proportion of *Labour* voters in the constituency (our parameter of interest).

Bayesian Inference - Example

We wish to specify a **prior distribution** for θ.

Since θ is a proportion (and hence must lie somewhere between 0 and 1) we choose a Beta distribution for the prior (since a Beta(α, β) distribution has a support of $(0, 1)$.

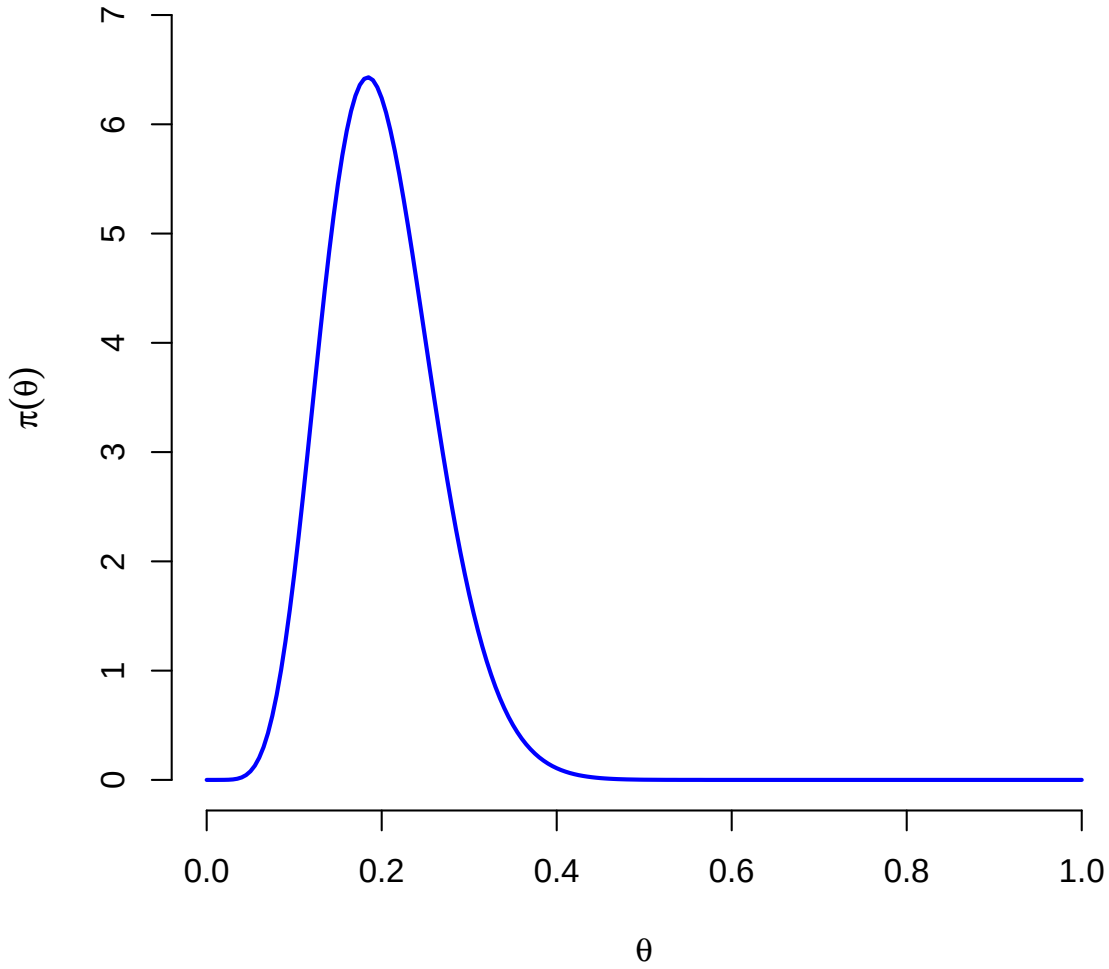
We specify the following prior distribution

$$\theta \sim \text{Beta}(8, 32)$$
$$\implies \pi(\theta) = \frac{\theta^7(1-\theta)^{31}}{B(8, 32)}.$$

This corresponds to a prior mean and standard deviation for θ of 0.2 and 0.062, respectively.

Bayesian Inference - Example

A plot of the density of the prior distribution for θ is shown below.



Bayesian Inference - Example

Now, suppose that in a random sample of 100 voters, 33 voters say that they will vote *Labour*.

Since each voter's response can be thought of as a Bernoulli random variable, then the sum of the 100 responses has a Bin$(100, \theta)$ distribution.

As such, we can write the likelihood function as

$$\mathcal{L}(\theta \mid \mathbf{x}) = \binom{100}{33} \theta^{33} (1-\theta)^{67}.$$

Bayesian Inference - Example

Recall that

$$\text{Posterior} \propto \text{ Prior} \times \text{Likelihood}.$$

Then, the posterior distribution for θ is

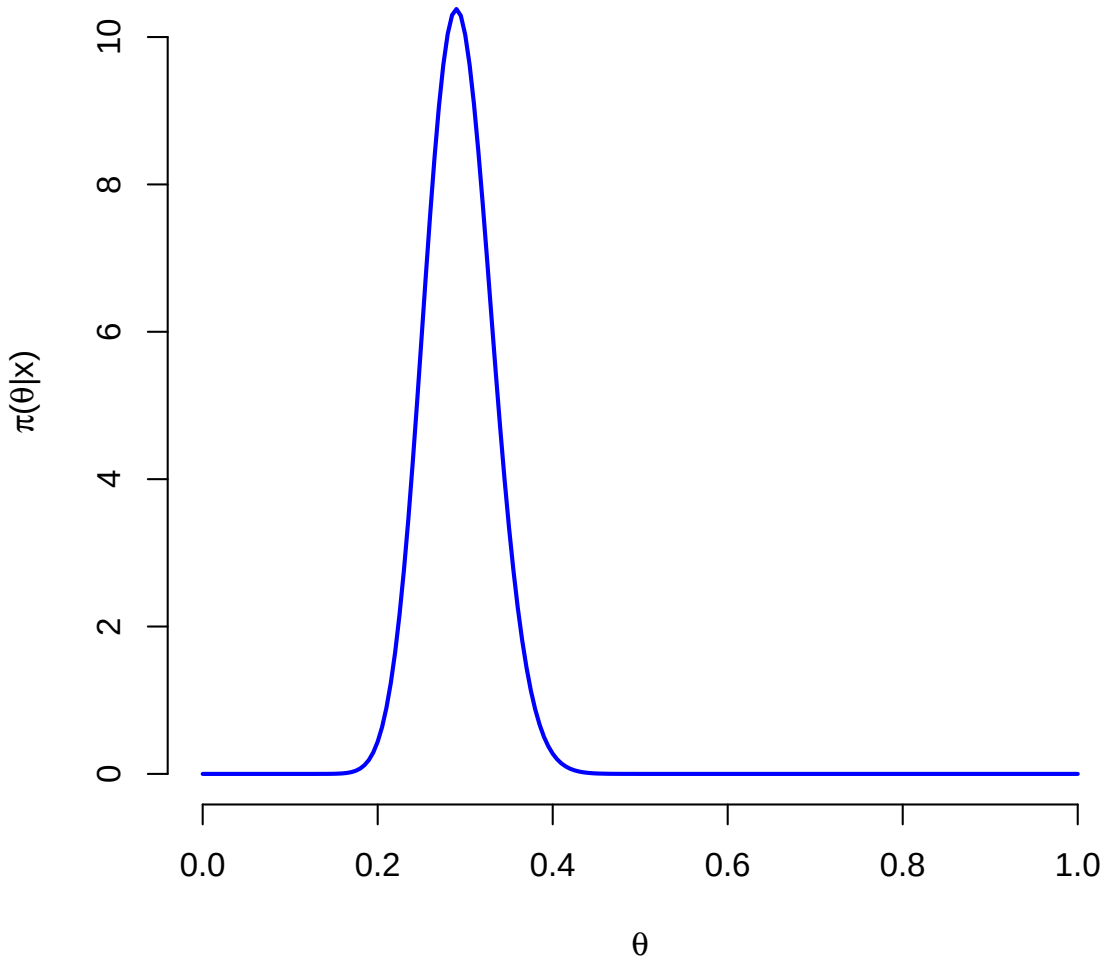
$$\begin{aligned}
\pi(\theta \mid \mathbf{x}) &\propto \pi(\theta) \times \mathcal{L}(\theta \mid \mathbf{x}) \\
&= \left[\frac{\theta^7(1-\theta)^{31}}{B(8,32)}\right] \times \left[\binom{100}{33}\theta^{33}(1-\theta)^{67}\right] \\
&= \frac{\binom{100}{33}}{B(8,32)}\theta^{40}(1-\theta)^{98} \\
&\propto \theta^{40}(1-\theta)^{98}
\end{aligned}$$

We see that $\pi(\theta \mid \mathbf{x})$ is proportional to the probability density function of a Beta$(41, 99)$ distribution. So the posterior distribution is

$$\theta \mid \mathbf{x} \sim \text{Beta}(41, 99)$$

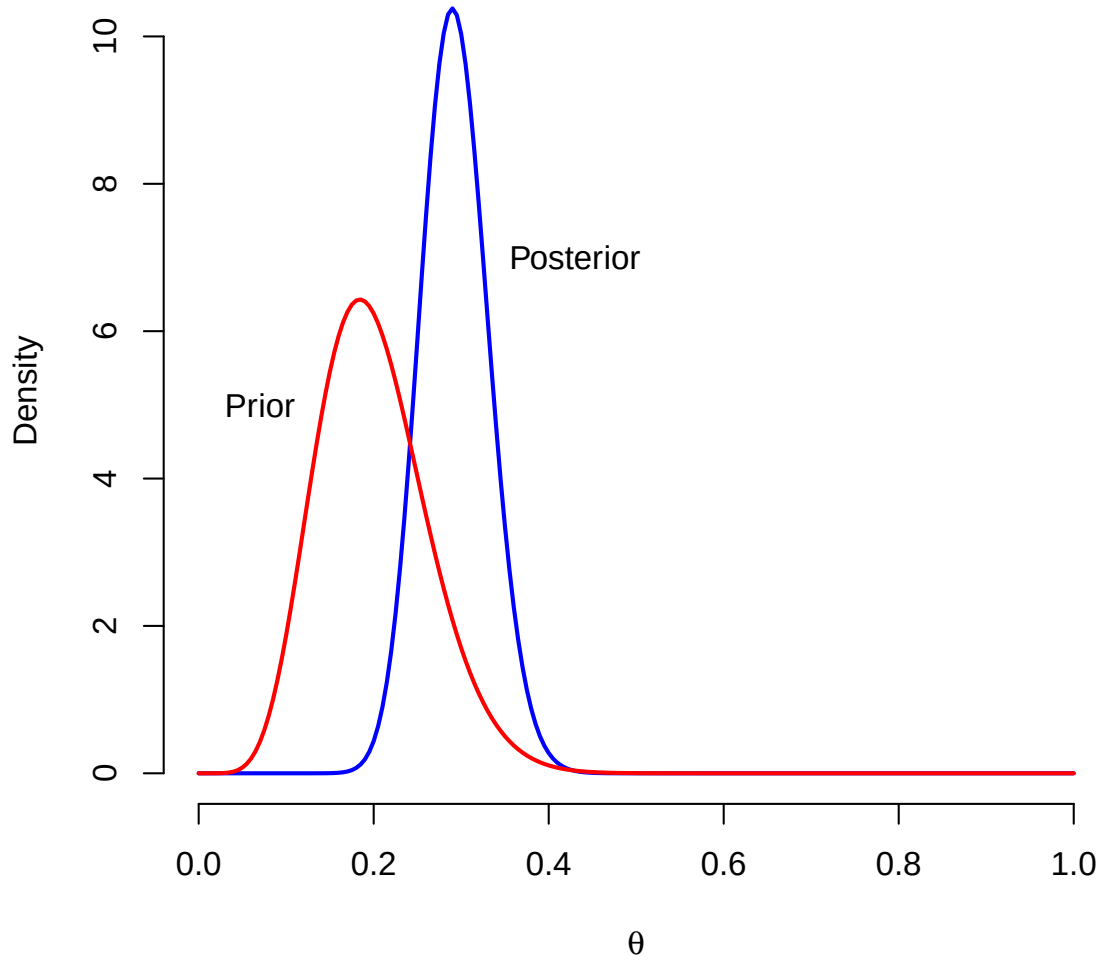
Bayesian Inference - Example

A plot of the density of the posterior distribution for θ is shown below.



Bayesian Inference - Example

Comparison with Prior.



Bayesian Inference - Example

The posterior distribution has mean $41/(41+99) \approx 0.29$. So the posterior distribution is shifted significantly from the prior (prior mean $= 0.2$). We could also compute 95% credible intervals etc. for θ and make other probabilistic statements.

Note that if more data became available then we could further update the posterior distribution.

The next posterior distribution for θ would be the same irrespective of whether we update the initial prior once and then again with a second set of data or if we combine all data into one sample and update our initial prior in one go.

Sufficiency in a Bayesian Context

If $T(\mathbf{X})$ is sufficient for θ then the joint density $f(\mathbf{x} \mid \theta)$ factorises as

$$f(\mathbf{x} \mid \theta) = f(\mathbf{x} \mid t) f(t \mid \theta)$$

Then, given a sample of data $\mathbf{x}$ and that $T(\mathbf{x}) = t$, the posterior distribution of θ is

$$\begin{aligned}
\pi(\theta \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta)\pi(\theta) \\
&= f(\mathbf{x} \mid t) f(t \mid \theta)\pi(\theta) \\
&= f(\mathbf{x} \mid t)\pi(\theta \mid t) f(t) \quad \text{because} \quad f(t \mid \theta) = \pi(\theta \mid t) f(t)/\pi(\theta) \\
&\propto \pi(\theta \mid t)
\end{aligned}$$

Hence, we see that the statistic $T(\mathbf{X})$ is sufficient for θ if and only if $\pi(\theta \mid \mathbf{x}) = \pi(\theta \mid t)$.

Statistical Decision Theory

We can think of statistical inference in terms of 'statistical decision theory'.

Broadly speaking, 'statistical decision theory' concerns the taking of a decision from a set of possibilities amidst uncertainty. For example: choosing an estimate for a parameter from a set of possibilities.

Statistical Decision Theory

In decision theory, we formulate a **general decision problem** in which the follow basic elements are contained.

- A **parameter space** Θ, where Θ contains all possible states of nature.
- An **unknown** true state of nature, θ, where $\theta \in \Theta$.
- A set $\mathcal{D}$ of all possible decisions/actions.
- A **loss function**, $L(\theta, d)$

$$L \colon \Theta \times \mathcal{D} \to \mathbb{R}.$$

which specifies the loss incurred if a decision d is taken when the true state of nature is θ.

Statistical Decision Theory

In statistical decision theory, it is assumed that a statistician may perform an experiment through which information may be obtained concerning the 'true state of nature'.

In particular, we assume that an observation (some data $\mathbf{x}$) is available that belongs to some sample space $\mathcal{X}$. The sample space $\mathcal{X}$ consists of all possible sets of data that could be sampled. We specify a **statistical decision function** $\delta(\mathbf{x})$ such that

$$\delta : \mathcal{X} \to \mathcal{D}.$$

Here, δ specifies the decision that should be taken having observed data $\mathbf{x}$.

The Loss Function

Now, as in our previous lecture, we could consider the set of random variables, $\mathbf{X}$, to be the data $\mathbf{x}$ prior to observation.

In this case, $\delta(\mathbf{X})$ (a function of the random variables $\mathbf{X}$) would be a random variable too.

As as result, we can consider the loss function associated with the decision $\delta(\mathbf{X})$ as a random variable $L(\theta, \mathbf{X})$.

Why is this useful. . . ?

The Loss Function

The answer is that we can then use the language of probability to consider the consequences of possible decisions given a set of data $\mathbf{X}$.

For example, we might calculate the average loss over all possible experimental outcomes, a measure known as a **risk function**, $R(\theta, \delta)$ where

$$R(\theta, \delta) = \mathbb{E}(L(\theta, \delta(\mathbf{X}))) = \int_{\mathcal{X}} L(\theta, \delta(\mathbf{x})) f(\mathbf{x}; \theta) \mathrm{d}\mathbf{x}.$$

Example: Hypothesis Testing

Suppose that we wish to perform a hypothesis test

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad \text{versus} \quad \mathrm{H}_1 : \theta \notin \Theta_0$$

where $\Theta_0 \subseteq \Theta$.

The test procedure runs as follows

1. Assume H_0 is true.
2. Given a set of data $\mathbf{x}$, where $f(\mathbf{x}; \theta)$ denotes the joint pdf/pmf of the random variable $\mathbf{X}$, compute a test statistic $t(\mathbf{x})$.
3. **Take a decision**: reject H_0 if $t(\mathbf{x}) > k$, retain H_0 if $t(\mathbf{x}) \leq k$ for some $k \in \mathbb{R}^+$.

So our possible states of nature are $\theta \in \Theta$ and the set of decisions is

$$\mathcal{D} = \{\text{Retain } \mathrm{H}_0, \text{Reject } \mathrm{H}_0\}.$$

Example: Hypothesis Testing

Here, the decision function is simply the test statistic $t(\mathbf{x})$ because this leads to the taking of the decision regarding the retention or rejection of the null hypothesis.

Based on the possible decisions, a loss function, $L(\theta, d)$ can be specified. This might look something like this

$$L(\theta, \text{Retain H}_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0; \\ \ell_{01} & \text{if } \theta \notin \Theta_0. \end{cases}$$

$$L(\theta, \text{Reject H}_0) = \begin{cases} \ell_{10} & \text{if } \theta \in \Theta_0; \\ 0 & \text{if } \theta \notin \Theta_0. \end{cases}$$

Here, ℓ_{01} denotes the loss incurred owing to a Type II error and ℓ_{10} denotes the loss incurred owing to a Type I error.

Criteria for Choosing a Decision Function

The structure of the decision space clearly depends on the type of inference that we wish to make.

For **point estimation** of the parameter θ, $\mathcal{D}$ contains the same set of points as the parameter space Θ and each decision corresponds to the choice of a value for the parameter θ. Thus $\delta(\mathbf{X})$ is an estimator of θ, with corresponding estimate $d = \delta(\mathbf{x})$.

For the construction of **confidence intervals**, D consists of ordered pairs of points from the parameter space.

For hypothesis testing, the decision space has just two elements, corresponding to rejection, or not, of the null hypothesis.

Choosing the Loss Function

An appropriate loss function also needs to be specified.

In point estimation of a single unknown parameter, θ, a quadratic loss function is often used in which $L(\theta, d)$ is taken to be proportional to the square of the estimation error, i.e. $L(\theta, d) \propto (d - \theta)^2$; this is known as **squared error loss**.

The corresponding risk function is then the familiar mean squared error (mse$(\delta(\mathbf{X}); \theta)$).

For a vector $\boldsymbol{\theta}$, a weighted sum of squares $(\mathbf{d} - \boldsymbol{\theta})^T \mathbf{W} (\mathbf{d} - \boldsymbol{\theta})$ is often used.

The Decision Function

If the true value of the parameter were known (say, $\theta = \theta_0$) then we would simply choose the decision d to minimise the loss $L(\theta_0, d)$.

For example, if we know $\theta = \theta_0$ then we can estimate θ perfectly so that the mean squared error is zero.

The inference problem arises because of our uncertainty about the true value of θ and the need to use the information provided by the sampled data $\mathbf{x}$.

There are then a number of possible approaches to determining $\delta(\mathbf{x})$.

We shall consider three important methods.

The Frequentist Approach

This approach is also referred to as the 'classical' or 'sampling theory' approach. The true parameter value θ is regarded as *fixed but unknown*. The approach involves concepts such unbiased estimators in point estimation and significance levels in hypothesis testing.

One main characteristic of the approach is that the properties of the chosen decision function (e.g. the point estimator, the critical region of the test) depend on the full sample space $\mathcal{X}$.

For example, if we say that a sample mean, $\bar{X}$, is unbiased for a parameter μ, then this means that the expected value of $\bar{X}$, taken over the whole sample space $\mathcal{X}$, is equal to μ.

The Likelihood Approach

This 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.

In other words, we concentrate on the sampling distribution $p(\mathbf{x}; \theta)$ as a function of θ over the whole parameter space Θ, but only at the single observed data point $\mathbf{x}$.

In order to emphasise this, we define the **likelihood function** $\mathcal{L}(\theta \mid \mathbf{x})$, or simply $\mathcal{L}(\theta)$, to be any function proportional to $p(\mathbf{x}; \theta)$, regarded as a function of θ.

The Bayesian Approach

In many ways this is close to the likelihood approach, as it also focuses on the observed data and ignores unobserved points in the sample space $\mathcal{X}$.

However, the unknown state of nature, the parameter θ, is regarded as random rather than fixed, and is assigned a probability distribution over the parameter space that is intended to reflect a degree of belief about the value of the parameter.

The ideas are straightforward – before any data are observed, θ is assumed to have a **prior distribution** $\pi(\theta)$; the data $\mathbf{x}$ then provide information about θ that can be used to update the prior beliefs about the true value of θ to give a **posterior distribution**, $\pi(\theta \mid \mathbf{x})$, using the likelihood function.

The Bayesian Approach

Recall the risk function $R(\theta, \delta)$. In a Bayesian analysis, since θ is considered a random variable, we can average the risk function over values of θ to yield a quantity known as the 'Bayes risk', $R(\delta)$ where

$$\begin{aligned} R(\delta) &= \int_{\Theta} R(\theta, \delta)\pi(\theta)\mathrm{d}\theta \\ &= \int_{\Theta}\int_{\mathcal{X}} L(\theta, \delta(\mathbf{x}))f(\mathbf{x};\theta)\pi(\theta)\mathrm{d}\mathbf{x}\mathrm{d}\theta \\ &= \int_{\mathcal{X}} \left\{ \int_{\Theta} L(\theta, \delta(\mathbf{x}))\pi(\theta \mid \mathbf{x})\mathrm{d}\theta \right\} f(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{X}} \mathbb{E}\left[L(\theta, \delta(\mathbf{x})) \mid \mathbf{x}\right] f(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned}$$

We see that the Bayes risk may be interpreted as the **posterior expected loss** on taking decision $\delta(\mathbf{x})$, given the observation $\mathbf{x}$. Note that $\mathbb{E}\left[L(\theta, \delta(\mathbf{x})) \mid \mathbf{x}\right]$ is taken with respect to the posterior distribution of θ given $\mathbf{x}$.

An Example

Suppose we define the random variable R as follows:

A fair coin is tossed and if it lands tails then $R = 0$.

If the coin lands heads then two balls are drawn from an urn where the urn contains a large number of balls and a proportion, θ, of the balls are red. Then, $R =$ Number of red balls drawn $+ 1$.

The probability mass function of R is given by

r	0	1	2	3
$\mathbb{P}(R = r)$	$\frac{1}{2}$	$\frac{1}{2}(1-\theta)^2$	$\theta(1-\theta)$	$\frac{1}{2}\theta^2$

Suppose that we wish to estimate θ, having observed the following ten values of r:

$$\mathbf{r} = (0, 0, 0, 0, 0, 0, 1, 1, 1, 2).$$

An Example: Frequentist Approach

The frequentist approach to the inference problem would focus on finding an unbiased estimator for θ. We'll start by computing $\mathbb{E}(R)$ and we see that

$$\begin{aligned}\mathbb{E}(R) &= \left(0 \times \frac{1}{2}\right) + \left(1 \times \frac{1}{2}(1-\theta)^2\right) + (2 \times \theta(1-\theta)) + \left(3 \times \frac{1}{2}\theta^2\right) \\ &= \frac{1}{2}(1+2\theta) = \frac{1}{2} + \theta\end{aligned}$$

So we see that $R - \frac{1}{2}$ is an unbiased estimator for θ. Our sample mean is

$$\bar{r} = \frac{1}{10}\left(6 \times 0 + 3 \times 1 + 1 \times 2\right) = \frac{1}{2}.$$

Substitution of $R = \bar{r}$ into the estimator for θ results in an estimate of 0.

Clearly, this is not a sensible estimate of θ - we know that the urn does not contain zero red balls!

An Example: Likelihood Approach

The likelihood function for θ is constructed using the joint probability density function of the observed data

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{r}) &= \left[\frac{1}{2}\right]^6 \left[\frac{1}{2}(1-\theta)^2\right]^3 [\theta(1-\theta)] \\ &= \left(\frac{1}{2}\right)^9 \theta(1-\theta)^7\end{aligned}$$

The likelihood function is maximised when $\theta = \frac{1}{8}$. So the maximum likelihood estimate is $\hat{\theta} = \frac{1}{8}$.

An Example: Bayesian Approach

In a Bayesian framework, we need a prior distribution for θ. Since $\theta \in (0, 1)$, we choose a $\mathcal{U}(0, 1)$ prior such that

$$\pi(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

We shall take the mean of the posterior distribution as our Bayes estimate of θ. Now

$$\begin{aligned} \pi(\theta \mid \mathbf{r}) &\propto \pi(\theta) \times \mathcal{L}(\theta; \mathbf{r}) \\ &\propto \theta(1-\theta)^7. \end{aligned}$$

So the posterior is proportional to the probability density function of a Beta$(2, 8)$ distribution.

If $\theta \mid \mathbf{r} \sim$ Beta$(2, 8)$, then the posterior mean (and Bayes estimate of θ) is 0.2.

Comment

It is clear that when the coin shows tails, the value of R does not provide any information on the proportion of balls in the urn, as the actual data values observed are not dependent on the value of θ.

The frequentist approach includes these data whereas the likelihood and Bayesian methods do not.

The Bayes estimator takes the prior information on θ into account and, in this particular example, this results in a Bayes estimate that is larger than the maximum likelihood estimate.

Learning Outcomes

- Understand the concept of **sufficiency**.
- Know how to use the factorisation criterion to identify sufficient statistics.
- Understand the concept of **Bayesian inference** and be able to find posterior distributions by combining priors and likelihood functions.
- Understand the concept of **loss functions** and **decision functions** and their use in statistical decision theory.
- Identify the differences between frequentist, likelihood and Bayesian approaches to decision problems.