# 2. Statistical Decision Theory

Decision is one of the most common human intellectual behaviors and mathematicians have spent over 300 years to analytically tackle this issue by establishing decision theory. Decision theory shall be able to answer the following questions:

- What is a decision?
- What is a good decision?
- How can we formally evaluate decisions?
- How can we formulate the decision problem confronting a decision maker?

## 2.1 Fundamentals

A decision problem involves a set $\mathcal{S}$ of states (that represents possible situations, affairs, etc.) and a set $\mathcal{X}$ of potential consequences of decisions. An *act* is considered as a mapping from the state space to the consequence set. That is, in each state $s \in \mathcal{S}$, an act $a$ delivers a well-defined result $a(s) \in \mathcal{X}$. The decision maker must rank acts without precise knowing current state of the world. In other words, an act is conducted with uncertainty.

It is a natural behavior for human beings to make a decision based on earlier experience that could often be modeled as statistics due to the uncertainty. *Statistical Decision Theory* is the mathematical framework to make decisions in the presence of statistical knowledge. Classical statistics uses sample information to directly infer parameter $\theta$. (Modern) Decision Theory makes the best decision by bringing sample information with relevant aspects of the problem, including

- Possible consequence of the decision introduced by Abraham Wald
- *A prior* information introduced by L.J. Savage 1961

The consequences of an act can often be ranked in terms of relative appeal, that is, some consequences are better than others. This is often numerically modeled via *utility function* $u$, which assigns a utility value $u(x) \in \Re$ for each consequence $x \in \mathcal{X}$. To model lacking knowledge about the state, we usually assume a probability distribution $p$ on $\mathcal{S}$, which can be obtained either from statistics (called *decision under risk* by Von Neumann and Morgenstern, 1944), or from a subjective probability

supplied by an agent through some methods.

The most common decision rule proceeds on the expected utility.

$$\mathbb{E}U(f) = \sum_{s\in\mathcal{S}} p(s)u[a(s)]$$

Different acts can be ranked based on preference of larger utility. As far as other criterion to deal with unknown current state of the world, we will introduce later.

We can therefore summarize the framework of statistical decision as follows. The unknown quantity $\theta$ affecting decision is commonly called the "state of nature". Let $\Theta$ denote the set of all possible states of nature. When the experiments are performed to obtain information regarding $\theta$, these experiments are typically designed such that the observations are distributed according to certain probability distribution of an unknown parameter $\theta$. We can therefore define the following.

- $\theta$: parameter
- $\Theta$: parameter space
- $a$: decision or action
- $\mathfrak{A}$: decision space or action sapce
- $L(\theta, a)$: loss function that is defined over all $(\theta, a) \in \Theta \times \mathfrak{A}$ and $L(\theta, a) > -\infty$

When a statistical investigation is conducted to obtain information regarding $\theta$, the outcome that is a random variable is denoted as $X$. A particular realization of $X$ is denoted as $x$. The set of possible outcomes is the *sample space* $\Omega$, and usually $\Omega \in \mathfrak{R}^n$.

The probability distribution of $X$ obviously depends on $\theta$, the unknown state of nature. $\forall A \subset \Omega$,

$$P_\theta(A) = \int_A f(x|\theta)dx \quad or \quad \sum_{x\in A} f(x|\theta)$$

For a given $\theta$, the expectation (over $X$) of a function $h(x)$ is

$$E_\theta[h(X)] = \int_\Omega h(x)f(x|\theta)dx = \int_\Omega h(x)dF^X(x|\,\theta)$$

It is straightforward to show (exercise 1) that

$$P_\theta(A) = \int_A dF^X(x|\theta)$$

Since *a prior* information regarding $\theta$ that might not be very precise, it is nature to describe *a prior* information in terms of probability distribution on $\Theta$, with $\pi(\theta)$ representing *a prior* density of $\theta$. If $B \subset \Theta$, we obtain

$$P(\theta \in B) = \int_B dF^{\pi}(\theta) = \int_B \pi(\theta) d\theta$$

### *2.1.1 Decision Rules and Risks*

Now, we are ready to mathematically define *decision* through a *decision rule* between sample space and decision space.

Definition 1: A non-randomized decision rule $\delta(x)$ is a function from $\mathcal{X}$ to $\mathfrak{A}$.
Definition 2: Two decision rules $\delta_1$ and $\delta_2$ are *equivalent* if $P_\theta[\delta_1(x) = \delta_2(x)] = 1, \forall\theta$.

As we mentioned that modern statistics introduce the concept of *risk* (or *cost*) associated with decision, we have the following.

Definition 3: The (average) risk function of a decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(x))]$$

Remark: For each $\theta$, $R(\theta, \delta)$ is thus the expected loss (over $X$) incurred in using $\delta(x)$.

It is vital to compare different decision rules.

Definition 4: $\delta_1$ is R-better than $\delta_2$ if $R(\theta, \delta_1) \leq R(\theta, \delta_2)\ \forall\theta \in \Theta$, and "<" holds for some $\theta$. $\delta_1$ is R-equivalent to $\delta_2$ if $R(\theta, \delta_1) = R(\theta, \delta_2)\ \forall\theta \in \Theta$.

Definition 5: $\delta$ is *admissible* if there exists no R-better decision rule. Otherwise, $\delta$ is *inadmissible*.

Of course, we only consider decision rule of finite risk.

Definiton 6: Let $\mathcal{D}$ denote the class of all decision rules $\delta$ for which $R(\theta, \delta) < \infty, \forall\theta$.

In many cases, non-randomized decision rules are not sufficed.

Definition: A *randomized decision rule* $\delta^*(x,\cdot)$ is $\forall x$ a probability distribution on $\mathfrak{A}$, if $x$ is observed, $\delta^*(x,A)$ is the probability that an action in $A$ $(A \subset \mathfrak{A})$ will be chosen. A *randomized action* $\delta^*(\cdot)$ is a probability distribution on $\mathfrak{A}$.

Example: Suppose $0 < p_i < 1$ and $\sum_{i=0}^{n} p_i = 1$, $\delta_i$ $(i = 1, \dots, n)$ are decision rules. $\delta = \sum_{i=0}^{n} p_i \delta_i$ is a randomized decision rule. ¶

Definition: The loss function $L(\theta, \delta^*(x,\cdot))$ of the randomized rule $\delta^*$ is defined as

$$L\big(\theta, \delta^*(x,\cdot)\big) = \mathbb{E}^{\delta^*(x,\cdot)}[L(\theta,a)]$$

The risk function of $\delta^*$ is

$$R(\theta,\delta^*) = \mathbb{E}_\theta^X\big[L\big(\theta,\delta^*(X,\cdot)\big)\big]$$

### *2.1.2 Decision Principles*

After having decision rules and risk functions, we still need a principle to make a decision. The most well known principle might be *Bayes* to make proper use of *a priori* information $\pi(\theta)$ regarding $\theta$, and loss function. In other words, for Bayes principle, *a priori* information and loss function must be available.

Proposition: A decision rule $\delta_1$ is preferred to rule $\delta_2$ if $\mathbb{E}^\pi[R(\theta,\delta_1)] < \mathbb{E}^\pi[R(\theta,\delta_2)]$. Therefore, the best decision rule according to *Bayes principle* is the one that minimizes (over all $\delta \in \mathcal{D}$)

$$r(\pi,\delta) = \mathbb{E}^\pi[R(\theta,\delta)]$$

Remark: $r(\pi,\delta)$ is called the Bayesian risk of $\delta$ (with respect to $\pi$). If a decision rule $\delta^\pi$ exists and minimizes $r(\pi,\delta)$ (over all $\delta \in \mathcal{D}$), then $\delta^\pi$ is called *Bayes rule*. $r(\pi) = r(\pi,\delta^\pi)$ is called the *Bayes risk* of $\pi$.

However, *a priori* information is not always available in decision. We therefore have to look another alternative, *minimax* principle. Supposing $\delta^* \in \mathcal{D}^*$, minimax principle considers $\sup_{\theta\in\Theta} R(\theta,\delta^*)$.

Proposition: A decision rule $\delta_1^*$ is preferred to rule $\delta_2^*$ if

$$\sup_\theta R(\theta,\delta_1^*) < \sup_\theta R(\theta,\delta_2^*)$$

Proposition: $\delta^{*M}$ is a *minimax decision rule* if it minimizes $\sup_\theta R(\theta,\delta^*)$ among all randomized rules in $\mathcal{D}^*$. That is,

$$\sup_{\theta\in\Theta} R(\theta,\delta^{*M}) = \inf_{\delta^*\in\mathcal{D}^*} \sup_{\theta\in\Theta} R(\theta,\delta^*)$$

Remark: If two decision problems have identical formal structures, then the same decision rule should be used in each problem, which is known as *invariance principle*.

Modern decision theory allows us to formulate problems more correctly and precisely. A good example is the commonly misused inference procedure in the hypothesis testing of a point null hypothesis, as the following example.

Example: A sample $X_1, \dots, X_n$ is to be taken from $\mathcal{N}(\theta, 1)$. We wish to conduct a test of $H_0: \theta = 0$ versus $H_1: \theta \neq 0$, at the significance interval 0.05. The usual test is to reject $H_0$ if $\sqrt{n}|\bar{x}| > 1.96$, where $\bar{x}$ denotes the sample mean.

However, in above example, it is not meaningful to state a point null hypothesis is rejected at a given level. From very beginning, we know the null hypothesis is almost certain not exactly true. A more realistic null hypothesis test would be $H_0: \theta \leq 10^{-3}$, for example.

Another decision-theoretic alternative has been considered. The following idea is very useful in digital communication theory. The first systematic development of *frequentist* ideas can be found in the work by J. Neyman and E. Pearson in 1967. The original motivation behind seems to produce measures that do not depending on $\theta$. It can be done by considering a procedure $\delta(x)$ and a criterion function $L(\theta, \delta(x))$, and then identify a number $R$ such that repeated use of $\delta$ would yield average long-run performance of $R$.

The conditional approach to statistics is concerned with reporting data-specific measure of accuracy. The major concern is the performance of $\delta(x)$ for actual data $x$ observed in a specific experiment.

Definition: For observed data $x$, the function $l(\theta) = f(x|\theta)$ is called the *likelihood function*.

Remark: The *likelihood principle* makes explicit the natural conditional idea that only the actual observed data $x$ should be relevant to the conclusions or evidence about $\theta$, while likelihood function plays the key role to facilitate likelihood principle.

Proposition: To make inferences or decisions about $\theta$ from the observed data $x$, all relevant information is contained in the likelihood function. Furthermore, two

likelihood functions contain the same about $\theta$ if they are proportional to each other.

To simplify the statistical problems, we wish to find a function of data that summarizes all available sample information about $\theta$, and we call this function as *sufficient statistics*.

Definition: Let $X$ be a random variable whose distribution depend on an unknown parameter $\theta$, and otherwise is known. A function $T$ of $X$ is the sufficient statistics for $\theta$, if the conditional distribution of $X$, given $T(X) = t$, is independent of $\theta$ (with probability 1).

We shall connect the concept of partition of the sample space with sufficient statistics.

Definition: If $T(X)$ is a statistic with range $\mathfrak{I}$ (i.e. $\mathfrak{I} = \{T(X): x \in \mathfrak{X}\}$), the partition of $\mathfrak{X}$ induced by $T$ is the collection of all sets of the form

$$\mathfrak{X}_t = \{x \in \mathfrak{X}: T(X) = t\}$$

for $t \in \mathfrak{I}$.

Remark: If $t_1 \neq t_2$, then $\mathfrak{X}_{t_1} \cap \mathfrak{X}_{t_2} = \emptyset$, and $\bigcup_{t\in\mathfrak{I}} \mathfrak{X}_t = \mathfrak{X}$. In other words, $\mathfrak{X}$ is partitioned into the disjoint sets $\mathfrak{X}_t$.

Definition: A *sufficient partition* of $\mathfrak{X}$ is a partition induced by a sufficient statistic $T$.

Theorem: Assuming $T$ to be a sufficient statistic for $\theta$, and $\delta_0^*(x,\cdot)$ to be any randomized decision rule in $\mathcal{D}^*$, then there exists a randomized decision rule $\delta_1^*(t,\cdot)$ Depending only on $T(X)$, which is $R$-equivalent to $\delta_0^*$.

Remark: Above is surely applied to non-randomized decision rule.
Remark: Likelihood Principle immediately implies that a sufficient statistic contains all the sample information regarding $\theta$.

### *2.1.3 Utility and Loss*

In formulating a statistical decision by evaluating the consequences of possible actions, we may encounter a problem, that is, the values of consequences may not have obvious scale(s) of measurement. Even there exists an obvious scale, and the scale might not be that meaningful. For example, 100 dollars for Alice might be quite

different from 100 dollars for Bob. To mathematically work on the "value", *utility theory* has been developed to assign appropriate numbers indicating such values.

All consequences of interest are called the *rewards* and denoted as $\mathcal{R}$, which is usually treated on real line but can also be non-numerical quantities. Since uncertainty exists for possible consequences, the results of actions can be modeled as probability distribution on $\mathcal{R}$. Let $\mathcal{P}$ denote the set of all such probability distributions. A real-valued function $U(r)$ can be constructed such that the value associated with a probability distribution $P \in \mathcal{P}$ would be given by the expected utility $E^P[U(r)]$. If such a function exists, it is called a *utility function*.

Definition: If $P_1$ and $P_2$ are in $\mathcal{P}$, then $P_1 \prec P_2$ stands for $P_2$ is preferred to $P_1$; $P_1 \asymp P_2$ means that $P_1$ is equivalent to $P_2$; and $P_1 \preccurlyeq P_2$ means that $P_1$ is not preferred to $P_2$.

We can use the following steps to construct $U$:

(i) We select two rewards $r_1$ and $r_2$ that are not equivalent. Assuming $r_1 < r_2$, let $U(r_1) = 0$ and $U(r_2) = 1$.

(ii) For a reward $r_3$ such that $r_1 < r_3 < r_2$, find $0 < \alpha < 1$ such that

$$r_3 \asymp P = \alpha\langle r_1\rangle + (1-\alpha)\langle r_2\rangle$$

(i.e. probability distribution giving probability $\alpha$ to $r_1$ and probability $1-\alpha$ to $r_2$) We can therefore define

$$U(r_3) = \alpha\, U(r_1) + (1-\alpha)U(r_2) = 1-\alpha$$

(iii) For a reward $r_3$ such that $r_1 > r_3$, find $\alpha$ such that

$$r_1 \asymp P = \alpha\langle r_3\rangle + (1-\alpha)\langle r_2\rangle$$

(iv) For a reward $r_3$ such that $r_3 > r_2$, find $\alpha$ such that

$$r_2 \asymp P = \alpha\langle r_1\rangle + (1-\alpha)\langle r_3\rangle$$

(v) Periodically check the construction process for consistency by comparing new combinations of rewards.

The reverse concept of utility is loss, which is more useful in most problems related to EE&CS, though having the equivalent mathematical structure. A common thinking to define the reverse operation may be negation, and other mathematical forms are possible.

### *2.1.4 Prior Information*

To this moment, we have assumed no information available about the true value of the parameter beyond that from data. However, an important element of modern decision problems is the *prior* information regarding the parameter (say, $\theta$) of interest, which is a probability distribution on $\Theta$. The classical concept is based on the frequency view. The theory of subject probability has been established when classical concept is insufficient, so that we may obtain $\pi(\theta)$. In the mean time, noninformative priors are also possible but it is easy to misuse.

There ways to determine *prior* information, while the most common ones include

- Maximum entropy
- Marginal distribution
- Prior selection via different approaches

Example: Assume $\Theta = \{\theta_1, \dots, \theta_n\}$. Maximum entropy yields $\pi(\theta_i) = \frac{1}{n}, i = 1, \dots, n$.

### *2.2 Statistical Decision Framework*

Given a statistical model, the information that we want to extract from data can be in various formats depending on our purpose. As a summary, we may have 4 common problems:

- Estimation, to deliver a "good guess" of an important parameter
- Testing, to know whether data support certain "specialness"
- Ranking, to give an order based on samples
- Prediction, given a vector $\mathbb{z}$, to say a random variable $Y$ of interest

Example (*Prediction*): We have a vector $\mathbb{z}$ representing data or observations, which can be used for prediction of a variable $Y$ of interest. This prediction rule is $\mu(\mathbb{z})$. Unfortunately, $\mu(\mathbb{z})$ is unknown. However, we have observations $(\mathbb{z}_1, Y_1), \dots, (\mathbb{z}_n, Y_n)$ to estimate $\mu(\cdot)$.

Remark: Statistical decision theory is essential to modern digital communication theory, which the transmitter essentially sends a binary information represented by one of the two possible waveforms, to the receiver through the channel. The receiver has to determine which one of the two possible waveforms is used, and thus to decide the binary information. It is a typical testing problem. However, to better judge the possible waveform, we may want to estimate the important parameter of

the possible waveforms.

Following earlier definitions, we start from the statistical model with an observation vector $\mathbb{X}$ whose distribution $P$ over a set $\mathcal{P}$ (where $\mathcal{P}$ is usually parameterized, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$). Now, we take the decision space or action space into scenario, which can be properly defined according to problems of interest.

The next step is more important to determine the loss function $L: \mathcal{P} \times \mathfrak{A} \to \Re^+$, which falls back to earlier representation $L(\theta, a)$ if $\mathcal{P}$ is parameterized.

Definition: To estimate a real-valued parameter either $\nu(P)$ or $q(\theta)$ when $\mathcal{P}$ is parameterized, we usually adopt loss function as
(*quadratic loss*) $L(\theta, a) = [q(\theta) - a]^2$
(*absolute value loss*) $L(\theta, a) = |q(\theta) - a|$
(*truncated quadratic loss*) $L(\theta, a) = \min\ \{[q(\theta) - a]^2, d^2\}$

Remark: Quadratic loss is most common, which is just like the squared distance. The truncated quadratic loss is related to *confidence interval loss*, which is useful in statistics and will be explored later. Although we introduce symmetric loss function here, loss function can be asymmetric and still useful.

Definition: We can further generalize to $D$-dimensional vector forms by treating $\mathbb{q} = \big(q_1(\vartheta), \dots, q_D(\vartheta)\big)$ and $\mathbb{a} = (a_1, \dots, a_D)$. Some common loss functions can be defined as

(*squared Euclidean distance*) $L(\vartheta, \mathbb{a}) = \frac{1}{D}\sum_{j=1}^{D}\big[q_j(\vartheta) - a_j\big]^2$

(*absolute distance*) $L(\vartheta, \mathbb{a}) = \frac{1}{D}\sum_{j=1}^{D}|q_j(\vartheta) - a_j|$

(*supreme distance*) $L(\vartheta, \mathbb{a}) = \max\ \{|q_1(\vartheta) - a_1|, \dots, |q_D(\vartheta) - a_D|\}$

Example: In the *prediction* problem, if we use $\rho(\cdot)$ as the predictor and observation $\mathbb{z}$ has marginal distribution $Q$, it is straightforward to consider

$$L(\theta, a) = \int [\mu(\mathbb{z}) - \rho(\mathbb{z})]^2 dQ(\mathbb{z})$$

as the expected squared error if $a$ is used, while $Q$ is the empirical distribution of $\mathbb{z}_j$ in the training $(\mathbb{z}_1, Y_1), \dots, (\mathbb{z}_n, Y_n)$. This leads to the common

$$L(\theta, a) = \frac{1}{D}\sum_{j=1}^{D}\left[\mu(\mathbb{z}_j) - \rho(\mathbb{z}_j)\right]^2$$

which is $\frac{1}{D}$ squared Euclidean distance between prediction vector and vector parameter.

Example: In the *testing*, we want to tell $\theta \in \Theta_0$ or $\theta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$. Or, equivalently $P \in \mathcal{P}_0$ or $P \in \mathcal{P}_1$. We may have the well known

(*0-1 loss*) $L(\theta, a) = \begin{cases} 0, \theta \in \Theta_a \text{ (correct decision)} \\ 1, otherwise \text{ (wrong decision)} \end{cases}$

which is useful in communication theory.

The data is a point $\mathbb{X} = \mathbb{x}$ in the sample space. A *decision rule* or *procedure* $\delta$ is any function from $\Omega$ to $\mathfrak{A}$ as Section 2.1.1.

Example (Estimation): To estimate a constant in the measurement, we may consider $\delta_1(x) = \bar{x}$ (sample mean) or $\delta_2(x) = \tilde{x}$ (sample median).

Example (Testing): In the well known *two-sample model*, $x_1, \dots, x_m$ denote $m$ subjects having a given disease with drug A, and $y_1, \dots, y_n$ denote $n$ other subjects having a given disease with drug B. If A is standard or placebo, $x_1, \dots, x_m$ are referred as *control observations* and $y_1, \dots, y_n$ are known as *treatment observations*. If $x_1, \dots, x_m$ are from the distribution $G(\mu, \sigma^2)$ and $y_1, \dots, y_n$ are from $G(\mu + \Delta, \sigma^2)$, we are asking the question whether the treatment effect parameter $\Delta = 0$ or not. Given an estimate $\hat{\sigma}$ of $\sigma$ from data, the decision rule is

$$\delta(\mathbb{x}, \mathbb{y}) = \begin{cases} 0, & \frac{|\bar{x} - \bar{y}|}{\hat{\sigma}} < \eta \\ 1, & \frac{|\bar{x} - \bar{y}|}{\hat{\sigma}} \geq \eta \end{cases}$$

where $\eta$ is called the *critical value* or *decision threshold*.

We currently know $\delta$ as the procedure, $L$ as the loss function, $\theta$ is the true value of the parameter, $\mathbb{X} = \mathbb{x}$ is the outcome from the experiment, then $L(P, \delta(\mathbb{x}))$ is the loss. However, $L(P, \delta(\mathbb{x}))$ is still unknown as $P$ is unknown. We further wish good properties over a wide range of $\mathbb{x}$, we therefore consider the *average* or *mean* loss over the entire sample space by treating $L(P, \delta(\mathbb{X}))$ as a random variable. We introduce the *risk function*

$$R(P,\delta) = \mathbb{E}_P[L(P,\delta(\mathbb{X}))]$$

as the measure of performance of the decision rule $\delta(\mathbb{X})$. $\forall\delta, R: \mathcal{P} \to \Re^+$ or $R: \Theta \to \Re^+$. $R(\cdot,\delta)$ is our a priori measure of performance of $\delta$.

Example (estimation): Let $\nu \triangleq \nu(P)$ be the real parameter to estimate and $\hat{\nu} \triangleq \hat{\nu}(X)$ be our estimator (i.e. from our decision rule). If we use squared loss, the resulting risk function is called *mean squared error* (MSE) of $\hat{\nu}$. That is,

$$\text{MSE}(\hat{\nu}) = R(P,\hat{\nu}) = \mathbb{E}_P[\hat{\nu}(X) - \nu(P)]^2$$

Definition: The *bias* of $\hat{\nu}$ is defined as

$$\mathscr{b}(\hat{\nu}) = \mathbb{E}(\hat{\nu}) - \nu$$

and can be considered as "long-run average error" of $\hat{\nu}$. If $\mathscr{b}(\hat{\nu}) = 0$, is it called *unbiased*.

Proposition: $\text{MSE}(\hat{\nu}) = [\mathscr{b}(\hat{\nu})]^2 + \text{Var}(\hat{\nu})$

*Proof:* $\hat{\nu} - \nu = [\hat{\nu} - \mathbb{E}(\hat{\nu})] + [\mathbb{E}(\hat{\nu}) - \nu]$ and immediately follows. ¶

Example (Testing): Continuing from earlier testing example, the decision rule between $\Delta = 0$ and $\Delta \neq 0$ only takes value 0 and 1. The risk is

$$R(\Delta,\delta) = L(\Delta,0)P[\delta(\mathbb{x},\mathbb{y}) = 0] + L(\Delta,1)P[\delta(\mathbb{x},\mathbb{y}) = 1]$$

Using 0-1 loss, we have

$$\begin{aligned} R(\Delta,\delta) &= P[\delta(\mathbb{x},\mathbb{y}) = 1] \text{ if } \Delta = 0 \\ &= P[\delta(\mathbb{x},\mathbb{y}) = 0] \text{ if } \Delta \neq 0 \end{aligned}$$

If $\mathfrak{X}$ and $\Theta$ denote the outcome space and parameter space, we are about to decide $\theta \in \Theta_0$ or $\theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset, \Theta_0 \cup \Theta_1 = \Theta$. A *test function* is a decision rule $\delta(\mathbb{X}) = 1$ on a set $C \subset \mathfrak{X}$ that is called the *critical region*, and $\delta(\mathbb{X}) = 0$ on $C^c$. That is, $\delta(\mathbb{X}) = 1_{\mathbb{X}\in C}$. Given $\theta \in \Theta_0$, $\delta(\mathbb{X}) = 1$ and we decide $\theta \in \Theta_1$, we refer such situation as *Type I error*. Given $\theta \in \Theta_1$, $\delta(\mathbb{X}) = 0$ and we decide $\theta \in \Theta_0$, we refer such situation as *Type II error*. Trying to identify a good test function is equivalent to finding the critical region with small probability of error. ¶

Proposition: In *Neyman-Pearson* framework of hypothesis testing, given a small bound (say, $\alpha > 0$) on the probability of Type I error, we minimize the probability of Type II error.

Remark: $\alpha$ is called the level of significance, and deciding $\Theta_1$ is known as "rejecting the hypothesis $H: \theta \in \Theta_0$ at level of significance $\alpha$".

It is still not quite enough for us to know up to this point. Considering an estimation problem as an example, it is natural for us to seek $v^{\mathrm{u}}(X)$ such that $\forall P$ of $X$

$$P[v^{\mathrm{u}}(X) \geq v] \geq 1 - \alpha$$

Then, such a $v^{\mathrm{u}}$ is called $(1-\alpha)$ upper confidence bound on $v$.

Definition: A *confidence interval* for , $\left[v^{\mathrm{l}}(X), v^{\mathrm{u}}(X)\right]$, is defined as

$$P[v^{l}(X) \leq v(P) \leq v^{u}(X)] \geq 1 - \alpha$$

Remark: As a summary, the statistical decision framework includes the following steps to treat:

(1) action space
(2) loss function
(3) decision procedure
(4) risk function
(5) confidence interval

*Comparison of Decision Procedures*

## *2.3 Prediction*

Recalling earlier defined prediction problem, suppose we know the joint distribution of a random vector $\mathbb{Z}$ and a random variable $Y$. We want to find a function $g$ on the range of $\mathbb{Z}$ such that $g(\mathbb{Z})$ (known as the *predictor*) is close to $Y$. A common way to define "close" is based on squared prediction error, $[g(\mathbb{Z}) - Y]^2$, when $g(\mathbb{Z})$ is used to predict $Y$. Since $Y$ is not known, we intend to use $\mathbb{E}[g(\mathbb{Z}) - Y]^2$, *mean squared prediction error* (MSPE). Of course, any distance (or metric) measure may be used and denoted by $d^2(g(\mathbb{Z}), Y)$.

Example (Linear Prediction): Let us consider a *finite duration impulse response* (FIR) discrete-time filter as the following figure with 3 functional blocks:

- $p$ unit-delay elements
- multipliers with weighting coefficients $w_1, w_2, \ldots, w_p$
- adders to sum over the delayed inputs to produce the output $\hat{x}[n]$

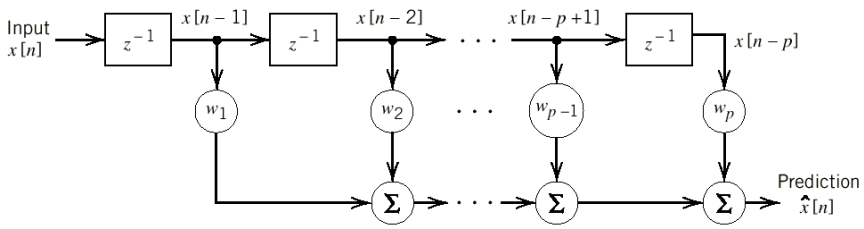

Figure: Linear Prediction Filter of Order *p*

More precisely, $\hat{x}[n]$, the *linear prediction* of the input is defined as the convolution sum

$$\hat{x}[n] = \sum_{k=1}^{p} w_k x[n-k]$$

The *prediction error*, is defined as

$$e[n] = x[n] - \hat{x}[n]$$

Our goal is to select $w_1, w_2, \ldots, w_p$ in order to minimize the performance index *J*, and we use *mean-square error* here.

$$J = E[e^2[n]]$$

The performance index is therefore

$$J = E[x^2[n]] - 2\sum_{k=1}^{p} w_k E[x[n]x[n-k]] + \sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k E[x[n-j]x[n-k]]$$

The input signal $x(t)$ is assumed from the sample function of a stationary process $X(t)$ of zero mean, and thus $E[x[n]] = 0 \ \forall n$. We define

$$\begin{aligned}\sigma_X^2 &= \text{variance of a sample of the process } X(t) \text{ at time } nT_s \\ &= E[x^2[n]] - (E[x[n]])^2 \\ &= E[x^2[n]]\end{aligned}$$

$$\begin{aligned}R_X(kT_s) &= \text{autocorrelation of the process } X(t) \text{ for a lag of } kT_s \\ &= R_X[k] \\ &= E[x[n]x[n-k]]\end{aligned}$$

*J* can be simplified as

$$J = \sigma_X^2 - 2\sum_{k=1}^{p} w_k R_X[k] + \sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k R_X[k-j]$$

We can reach the necessary condition of optimality by differentiating filter coefficients, then we can get the famous *Wiener-Hopf* equation for linear prediction.

$$\sum_{j=1}^{p} w_j R_X[k-j] = R_X[k] = R_X[-k], \qquad k = 1,2,\ldots,p$$

Theorem: (Wiener-Hopf equation in matrix form) Let

$\mathbf{w_o}$ = $p$-by-1 optimum coefficient vector

$= [w_1, w_2, \ldots, w_p]^T$

$\mathbf{r_X}$ = $p$-by-1 autocorrelation vector

$= [R_X[1], R_X[2], \ldots, R_X[p]]^T$

$\mathbf{R_X}$ = $p$-by-$p$ autocorrelation matrix

$$= \begin{bmatrix} R_X[0] & R_X[1] & \cdots & R_X[p-1] \\ R_X[1] & R_X[0] & \cdots & R_X[p-2] \\ \vdots & \vdots & & \vdots \\ R_X[p-1] & R_X[p-2] & \cdots & R_X[0] \end{bmatrix}$$

Then,

$$\mathbf{R_X w}_o = \mathbf{r_X}$$

## *2.4 Sufficiency*

Definition: A statistic $T(\mathbb{X})$ is called *sufficient* for $P \in \mathcal{P}$ or parameter $\theta$ if the conditional distribution of $\mathbb{X}$ given $T(\mathbb{X}) = t$ does not involve $\theta$.

Remark: Once a sufficient statistic $T$ is known, the sample $\mathbb{X} = (X_1, \ldots, X_n)$ does not contain any further information about $\theta$ or equivalently $P$, given $\mathcal{P}$ is valid.

Remark: If $T_1$ and $T_2$ are two statistics such that $T_1(x) = T_1(y)$ if and only if $T_2(x) = T_2(y)$, then $T_1$ and $T_2$ provide the same information and achieve the same reduction of the data. $T_1$ and $T_2$ are called equivalent.

Theorem (Factorization Theorem): In a regular model, a statistic $T(\mathbb{X})$ with range $\mathcal{T}$ is sufficient for $\theta$, if and only if, there exists a function $g(t, \theta)$ defined for $t \in \mathcal{T}$ $\theta \in \Theta$ and a function $h$ defined on $\mathfrak{X}$ such that

$$p(\mathbb{x}, \theta) = g(T(\mathbb{x}), \theta) h(\mathbb{x})$$

$\forall \mathbb{x} \in \mathfrak{X}, \theta \in \Theta$.

*Proof*: The complete proof can be found in a well known book by Lehmann (1977). ¶

Remark: Outcome space $\mathfrak{X}$ is generally just sample space $\Omega$.

Example: Suppose i.i.d. random variables $X_1, \ldots, X_n \sim G(\mu, \sigma^2)$ and both mean and variance are unknown. Let $\theta = (\mu, \sigma^2)$.

$$p(x_1, \ldots, x_n, \theta) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2}$$

We note that $p(x_1, \ldots, x_n, \theta)$ is a function of $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ and $\theta$ only.

Applying above theorem, we can conclude

$$T(X_1, \dots, X_n) = \left( \sum_{i=1}^{n} X_i , \sum_{i=1}^{n} X_i^2 \right)$$

to be sufficient for $\theta$. Another equivalent sufficient statistic that is frequently used is

$$T_{eq}(X_1, \dots, X_n) = \left( \frac{1}{n} \sum_{i=1}^{n} X_i , \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is called the *sample mean* and the second term is the *sample variance*. ¶

Sufficiency can be clearly described in the statistical decision theory. Specifically, if $T(\mathbb{X})$ is sufficient, $\forall \delta(\mathbb{x})$, we can find a randomized decision rule $\delta^*\big(T(\mathbb{X})\big)$ depending only on $T(\mathbb{X})$, and $R(\theta, \delta) = R(\theta, \delta^*), \forall \theta$. By randomization, $\delta^*\big(T(\mathbb{X})\big)$ can be generated from the value $t$ of $T(\mathbb{X})$ and a random mechanism not depending on $\theta$.

Definition: $T(\mathbb{X})$ is *Bayes sufficient* for $\Pi$ if the *posterior* distribution of $\theta$ given $\mathbb{X} = \mathbb{x}$ is the same as the posterior (conditional) distribution of $\theta$ given $T(\mathbb{X}) = T(\mathbb{x}), \forall \mathbb{x}$.

Theorem (Kolmogorov): If $T(\mathbb{X})$ is sufficient for $\theta$, it is Bayes sufficient for every $\Pi$.

If $T(\mathbb{X})$ and $S(\mathbb{X})$ are sufficient, but $T(\mathbb{X})$ provides a greater reduction of data. We say $T(\mathbb{X})$ to be *minimally sufficient*, if it is sufficient and provides a greater reduction of data than any other sufficient statistic $S(\mathbb{X})$. By factorization theorem, we can find a transform $\varsigma$ such that $T(\mathbb{X}) = \varsigma[S(\mathbb{X})]$. On the contrary side of minimally sufficient data, general data may have the *irrelevant* part, which is not useful for us neither to postulate nor to infer useful information.

Since we can use $p(\mathbb{x}, \theta)$ for different values of $\theta$ and use factorization theorem to construct the minimally sufficient statistic. We can define the likelihood function $\mathcal{L}$ for a given observed data vector $\mathbb{x}$ as

$$\mathcal{L}_{\mathbb{x}}(\theta) = p(\mathbb{x}, \theta), \theta \in \Theta$$

Consequently, $\mathcal{L}_{\mathbb{x}}$ is a mapping from the outcome space $\mathfrak{X}$ (or sample space $\Omega$) to the class $\mathcal{T}$ of functions $\{\theta \to p(\mathbb{x}, \theta) : \mathbb{x} \in \mathfrak{X}\}$. For a given $\theta$, $\mathcal{L}_{\mathbb{x}}(\theta)$ represents the probability (density) of observing $\mathbb{x}$. Through the Bayes Theorem,

$$(\text{posterior}) \propto (\text{prior}) \propto (\text{likelihood})$$

*Exercises:*

1. Please prove

$$P_\theta(A) = \int_A dF^X(x|\theta)$$

2. For a real-numbered monotonically increasing sequence $\{a_n\}, n = 1, .2, \ldots,$ please find an example such that $\max_n a_n \neq \sup_n a_n$.
3. Consider a parameter space consisting of two pints $\theta_1$ and $\theta_2$. For given $\theta$, an experiment leads to a random variable $X$ whose frequency function $p(x|\theta)$ is given by

| $\theta \backslash x$ | 0 | 1 |
|---|---|---|
| $\theta_1$ | 0.8 | 0.2 |
| $\theta_2$ | 0.3 | 0.7 |

Let $\pi$ be the prior frequency function of $\theta$ defined by $\pi(\theta_1) = \frac{1}{2}, \pi(\theta_2) = \frac{1}{2}$.

(a) Find the posterior frequency function $\pi(\theta|x)$.

(b) Suppose $X_1, \ldots, X_n$ are independent with frequency function $p(x|\theta)$. Find the posterior $\pi(\theta|x_1, \ldots, x_n)$. Please note that it depends only on $\sum_{i=1}^n x_i$.

(c) Repeat (b) for $\pi(\theta_1) = \frac{1}{3}, \pi(\theta_2) = \frac{2}{3}$.

4. Suppose $p(x|\theta) = e^{-(x-\theta)}, 0 < \theta < x$ and $\pi(\theta) = 2e^{-2\theta}, \theta > 0$. Please find the posterior density $\pi(\theta|x)$.
5. Let $\overline{X_b}$ and $\widetilde{X_b}$ denote the sample mean and the sample median of the sample $X_1 - b, \ldots, X_n - b$. If the parameters of interest are the population mean and median of $X_i - b$ respectively, please show that $\text{MSE}(\overline{X_b})$ and $\text{MSE}(\widetilde{X_b})$ are the same for all $b$ (i.e. invariant with respective to shift).
6. An urn contains $N$ red balls and $N$ green balls. $2n$ $(n < N)$ balls are drawn at random without replacement. $R_n$ denotes the number of red balls in the first $n$ draws and $R_t$ denotes the total number of red balls drawn.

(a) Please find the best predictor of $R_t$ given $R_n$.

(b) Please find the best linear predictor of $R_t$ given $R_n$.

(c) Please find the MSPE for (a).

7. $X_1, X_2 \sim \mathcal{N}(0,1)$ and are independent. Let $Z = X_1^2 + X_2^2$ and $Y = X_1$.

(a) Is *Z* useful to predict *Y*?

(b) Is *Y* useful to predict *Z*?

8. Suppose $X_1, \ldots, X_n$ is a sample from a population one of the following density functions. For each case, please find a sufficient statistics for fixed $\theta, a$.

(a) $p(x,\theta) = \theta x^{\theta-1},\ 0 < x < 1, \theta > 0$. This is obviously Beta distribution $\beta(\theta, 1)$.

(b) $p(x,\theta) = \theta a x^{a-1} e^{-\theta x^a}, x > 0, \theta > 0, a > 0$. This is known as the *Weibull* distribution.

(c) $p(x,\theta) = \frac{\theta a^\theta}{x^\theta}, x > a, \theta > 0, a > 0$. Thi is known as the *Pareto* distribution.

9.