BIOSTAT 802: Advanced Statistical Inference II
Statistical Decision Theory: Part I

Peter XK Song

1 Basic Concepts of Statistical Decision

In late 1940's, Dr. Abraham Wald proposed a new point of view for statistical analysis that regarded statistical inference as a kind of "game" between human and the nature. This point of view, which was later termed as *statistical decision theory*, has made some significant positive effects on facilitating the development of a unified theory on various statistical methodologies. From the perspective of Wald's decision theory, any statistical problem (e.g. parameter estimation, hypothesis testing, confidence interval estimation, Bayesian statistics) may be thought of as a special kind of statistical decision. In effect, this novel point of view did make a remarkable impact on the development of statistical theory and methods in decades after. However, to be clear, this role of the statistical decision theory is primarily to extend the spectrum of statistical problems, rather than to provide analytic tools to problem solving. This section focuses on a systematic introduction to basic concepts of the statistical decision theory.

1.1 Three Elements in the Statistical Decision Theory

In order to estimate a parameter, we give an estimator that is, mathematically speaking, a function of data. Likewise, in order to perform a hypothesis test, we provide a test statistic whose distribution is manageable to establish a certain rejection rule, which is also a function of data. In order to make a recommendation, we establish an evaluation machinery based on some training data. Either an estimator, or a test statistic, or a recommendation, is a solution to a statistical problem. Generally speaking, a solution to a statistical problem may be regarded as a *statistical decision function (or decision rule)*. To formally define such an important concept of statistical decision function, in what follows we need to first build up a framework of statistical decision theory that contains basic elements relevant to a statistical

problem. This pertains to the so-called set of three elements, including *sample space and distribution family, action space and loss function.*

1.1.1 Sample Space and Distribution Family

Samples or observations, or data, are the inforamtion required to derive a statistical solution to a statistical problem. Theoretically speaking, a sample may be regarded as a set of random variables (or random vectors) that follow a certain distribution, part of which may be unknown. To be clear in this section, we use X to denote a sample of one-dimension or multi-dimenison while x denotes a realized observation of X. When there is no confusion, sometimes X and x are used exchangeably in this section.

Denote the collection of all possible values of X by $\mathcal{X}$. For the sake of mathematical convenience, sometimes, $\mathcal{X}$ may be extended to include some impossible values of X. For example, in a problem where X can only take nonnegative values, for convenience, one may assume $X \in R^1 = (-\infty, \infty)$, where the range of $(-\infty, 0)$ of impossible values are included to enlarge the domain of X. From now on, R^n denotes the n-dimensional Euclidean space.

In order to define a probability distribution of X, we need to construct a σ-field that lists subsets of $\mathcal{X}$. The corresponding σ-field, no matter how it may be constructed, is denoted by $\mathcal{B}_x$. Thus, the pair $(\mathcal{X}, \mathcal{B}_x)$ is a measurable space and called the *sample space* with respect to a statistical problem of interest. In most of statistical problems, $\mathcal{X}$ is specified as a finite-dimensional Euclidean space R^n or a non-empty Borel subset of R^n. In this case, $\mathcal{B}_x$ is given by the σ-field generated by all possible Borel subsets of $\mathcal{X}$, and the resulting sample space $(\mathcal{X}, \mathcal{B}_x)$ is said to be *Euclidean*. Often, we state that random variable X has or is defined on a sample space $(\mathcal{X}, \mathcal{B}_x)$.

As mentioned above, sample X follows a certain probability distribution, at least part of which may be unknown. A purpose of acquiring an observation x is to hopefully learn the unknown part of the distribution from the observed data. Typically, we formulate a statistical problem by assuming that the distribution of X belongs to a certain family of distributions, denoted by $\mathcal{P}$. In some statistical problems, the construct of $\mathcal{P}$ is assumed to be known. Thus, a natural question raises:

when we face to a statistical problem, how would we specify $\mathcal{P}$?

Although it is nearly impossible to answer this important question from a theoretical point of view, it is possible to do so from a practical point of view when a specific applied problem is fixed for investigation. Generally speaking, setting up $\mathcal{P}$ requires relevant knowledge of substantive sciences and cumulative experiences from the study of previously similar problems. Sometimes, specifying $\mathcal{P}$ may be purely based on a hypothesis or based on consideration of mathematical simplicity. For example, a family of normal distributions is frequently used in a statistical problem.

In order to pinpoint a specific distribution in a family $\mathcal{P}$, we often introduce a label attached to each distribution; for example, a symbol θ. Consequently, family $\mathcal{P}$ may be rewritten as $\{P_\theta, \theta \in \Theta\}$, where Θ is a set of labels; and when θ varies over the entire set Θ, P_θ visits every member of $\mathcal{P}$. Such label θ is conventionally termed as *distribution parameter*. From now on, we assume that different values of the distribution parameter θ correspond to different distributions. Moreover, set Θ that collects all possible values of θ is called the *parameter space*. In this course, we consider the case that Θ is a finite-dimensional Euclidean space or a non-empty Borel subset. For convenience, sometimes we need to introduce a σ-filed with respect to Θ, i.e. a σ-field generated by all Borel subsets of Θ; in other words, parameter (measurable) space $(\Theta, \mathcal{B}_\Theta)$ is Euclidean.

Commonly we consider a scenario where the distribution family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ satisfies the following condition: there exists a σ-finite measure μ on $\mathcal{B}_x$ such that $\mathcal{P}_\theta \ll \mu$ (or $\mathcal{P} \ll \mu$) for all $\theta \in \Theta$. Consequently we may rewrite $\mathcal{P}$ in the form of $\{p_\theta d\mu, \theta \in \Theta\}$, where p_θ is a Radom-Nikodym derivative of P_θ with respect to the dominated measure μ. In this course, we consider two major types of measure μ. First, μ is the Lebesgue measure, and second, μ is the counting measure on a countable set that is defined as the cardinality of a set A, or $\mu(A) = \text{card(A)}$.

Joining distribution family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ and sample space $(\mathcal{X}, \mathcal{B}_x)$ leads to a triplet $(\mathcal{X}, \mathcal{B}_x, \mathcal{P}) = (\mathcal{X}, \mathcal{B}_x, P_\theta, \theta \in \Theta)$, which is named as the *probability space*, and is sometimes also called the sample space.

Let us look at a few examples.

Example 1 To estimate weight of an object, b, n independent measurements are collected, and denoted by $X_1, \ldots, X_n$. Suppose each measurement error follows a normal distribution

$N(0, \sigma^2)$, where the parameter $\sigma > 0$ is unknown. In this problem, the sample is an n-dimensional vector $X = (X_1, \ldots, X_n)$, and the sample space is $(R^n, \mathcal{B}^n)$, where $\mathcal{B}^n$ is the σ-field consisting of all Borel subsets of R^n. The parameter space is $\Theta = \{\theta : \theta = (b, \sigma), -\infty < b < \infty, \sigma > 0\}$, and the distribution family is $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ where

$$dP_\theta(x) = (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - b)^2\right\} dx_1 \cdots dx_n.$$

Note that here impossible values are included in both $\mathcal{X}$ and Θ for mathematical convenience.

Example 2 To estimate the rate of plates with flaws sold in a shop (denoted by θ), an inspector randomly selects n plates out of N plates. Let X denotes the number of plates with flaws. When $n/N \approx 0$ or sampling with replacement is conducted, X approximately follows a binomial distribution $B(n, \theta)$. In this example, the sample space is naturally $\mathcal{X} = \{0, 1, \ldots, n\}$ and $\mathcal{B}_x$ is the collection of all subsets of $\mathcal{X}$. The parameter space is $\Theta = \{\theta : 0 \leq \theta \leq 1\}$ and the distribution family is

$$dP_\theta(x) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} d\mu(x), \ 0 \leq \theta \leq 1,$$

where μ is the counting measure on $\mathcal{X}$.

Example 3 Consider Example 1. Now assume that the n measurement errors are independent and identically distributed (iid), and that each measurement error follows a distribution with mean zero (yes, that is all we know). The sample space remains the same as that given in Example 1. The distribution family $\mathcal{P}$ is constructed as follows: $P = F \times F \times \cdots \times F$ (n-tuple orthogonal product), where F is any 1-dimensional distribution function with mean zero. In this case, one may use F as a label, and write $P = P_F$. Thus, the parameter space is

$$\Theta = \{F : F \text{ is a 1-dim distribution function satisfying } \int_{-\infty}^{\infty} |x| dF(x) < \infty \text{ and } \int_{-\infty}^{\infty} x dF(x) = 0\}.$$

There are two aspects in that this example differs from Example 1. First, the parameter space is no longer Euclidean; and second, for each individual distribution P_F in the distribution family $\mathcal{P}$, there is no closed-form expression to analytically express the relationship between P_F and F. In the literature, this case is usually referred to as a *nonparametric problem*,

whereas the other two above as *parametric problems*. Sometimes, such difference is vague and hard to be defined precisely, and in most of cases, making such difference clear may be really unnecessary.

1.1.2 Action Space

Action space is also called decision space. For a specific statistical problem, depending on the nature of the problem, its solution may be a number in the setting of point estimation, an interval in the setting of confidence interval estimation, or a decision of rejection or acceptance in the setting of hypothesis test, or a recommendation of favor or unfavor or withhold in a recommender system. The problem of statistical decision can be formulated as a general process of seeking answers to various statistical problems, and the key is decision. Because every decision (e.g. rejecting or accepting a hypothesis) leads to an action, an answer to a problem of statistical decision is commonly thought of as an action. For a given problem of statistical decision, which action is taken depends on multiple factors, such as the nature of problem, the characteristics of data collected, and the optimality criterion. Interestingly, in many scenarios, it is possible to know *a prior* all possible actions for a given statistical problem. Thus, the set of all possible actions is referred to the *action space*, denoted by A. For the mathematical rigor, one may extend A in a suitable way, so that the resulting expansion and the related σ-field $\mathcal{B}_A$ generated (typically) by all possible Borel subsets of the expanded A can form a measurable space $(A, \mathcal{B}_A)$.

Example 4 Consider Example 1. If point estimation of mean weight b is of interest, then the action space $A = \{a : a \in R^1\}$. If an interval estimation of b is of interest, then the action space $A = \{a : a = [a_1, a_2], a_1 < a_2, (a_1, a_2) \in R^2\}$. If simultaneous point estimation of b and σ is of interest, then the action space is $A = \{a = (a_1, a_2) : -\infty < a_1 < \infty, a_2 > 0\}$.

Example 5 Consider k many normal distributions $N(\beta_i, \sigma^2), i = 1, \ldots, k$. The problem is to estimate the ascent ordering of the k population means, $\beta_1, \ldots, \beta_k$, based on certain samples drawn from these populations. For the simplicity, assume there are no ties among the k population means. In this case, each action pertains to a permutation, $(i_1, \ldots, i_k)$, of $1, 2, \ldots, k$, where mean β_{i_1} is the smallest, mean β_{i_2} is the second smallest, and so on,

and mean β_{i_k} is the largest. Thus, the action space A contains $k!$ many elements, each corresponding to a permutation of $1, \ldots, k$.

It is worth pointing out that for a problem of hypothesis test, the action space is rather simple, consisting of only two actions: rejection and acceptance.

Likewise, for a problem of recommendation, the action space consists of pre-fixed kinds of recommendations. For example, in a paper review setting, a reviwer makes a recommendation to the journal editor by takeing one item from a set of recommendation options, typically consisting of acceptance, minor revision, major revision, rejection but encourage to resubmission, and rejection.

1.1.3 Loss Function

For a problem of statistical decision, any action taken incurs a good or bad consequence, which may be measured as loss in dollars or as such. For example, in Example 2, the inspector needs to make a decision if the shop would accept the N plates or not. If the rate of plates with flaws θ is small, the decision of accepting the plates is associated with a low loss; otherwise, it would be a disaster.

The statistical decision theory assumes that essentially all steps in the decision-making, including the ultimate evaluation of loss or benefit, can be quantified numerically. If we use loss as the ultimate objective to evaluate the goodness of an action, a reasonable one should satisfy the following obvious properties: loss is always nonnegative, and the closer an action is to the correct one the smaller loss incurs. This gives rise to an important concept of loss function. A loss function $L(\theta, a)$ is a nonnegative function defined on $\Theta \times A$. It may be interpreted as: the loss of an action a is $L(\theta, a)$ when the true parameter is θ.

For example, in Example 1, for a problem of point estimation for mean parameter b, we may adopt a loss of function of the following form:

$$L(\theta, a) = (b - a)^2,$$

where $\theta = (b, \sigma)$. This is the famous *squared error loss* function, which plays a central role in the classical theory of point estimation. A more general form may be given by

$$L(\theta, a) = c(\theta)w(b - a)$$

where $c(\theta) > 0$ and $w(x)$ is an even function and non-decreasing for $x \geq 0$. An important choice of $w(x)$ is from the family of convex functions, such as the *absolute error loss* function, i.e. $w(x) = |x|$.

In the problem of interval estimation, say for b in Example 1, we consider the loss from two aspects for an action $a = [a_1, a_2]$. One concerns whether the interval covers the mean parameter b or not, and the other pertains to the length of the interval. One possible specification may be given by

$$L(\theta, a) = m\{1 - I_{[a_1,a_2]}(b)\} + (a_2 - a_1),$$

where $m > 0$ is a constant reflective to the amount of loss when the interval does not cover b. Another choice might be

$$L(\theta, a) = |a_1 - b| + |a_2 - b|.$$

In the problem of hypothesis test, since there are only two actions in the action space, the specification of loss function may be relatively straightforward. For example, one can take a 0-1 loss function: the loss is zero if a right decision is taken; otherwise the loss is 1. This loss function is actually used in the famous Neyman-Pearson Lemma. Technically, any specified loss function $L(.,.)$ needs to satisfy, mathematically, the condition of measurability. That is, $L(\theta, a)$ is measurable on the σ-field $\mathcal{B}_A$ when θ is fixed, namely

$$L(\cdot; a) : (A, \mathcal{B}_A) \to (R, \mathcal{B}_R);$$

otherwise, in general, $L(\theta, a)$ is measurable on the σ-field $\mathcal{B}_\Theta \times \mathcal{B}_A$, namely

$$L(\theta; a) : (\Theta, \mathcal{B}_\Theta) \times (A, \mathcal{B}_A) \to (R, \mathcal{B}_R).$$

An obvious weakness of the statistical decision theory is the ambiguity and arbitrariness of loss function specification. First, it is impractical to assume that the loss of any action in all problems can be quantified numerically. For example, it is extremely hard to quantify entical and fairness principles in clinical studies. Second, if numerical quantification of loss is acceptable, determining a suitable form of loss function is subjective. In theory, most of time people use simplicity as the criterion to choose a loss function.

1.2 Decision Function and Risk Function

1.2.1 Decision Function

As mentioned above, a decision function (or a decision rule) provides a solution or an answer to a statistical problem. As a solution, it defines an action, according to available observed sample x. In this sense, a decision function is a mapping from the sample space to the action space. Suppose that one uses a decision function δ; then when a sample x is obtained, an action $\delta(x)$ is taken. Technically, the measurability of a decision function $\delta(x)$ is required. That is, a valid decision function δ should be a measurable transformation from the sample space $(\mathcal{X}, \mathcal{B}_x)$ to the action space $(A, \mathcal{B}_A)$. That is,

$$\delta^{-1}(D) = \{x \in \mathcal{X} : \delta(x) \in D\} \in \mathcal{B}_x, \ \forall D \in \mathcal{B}_A.$$

The above decision function is referred to as a *deterministic* decision function. By δ being deterministic it means that the resulting action $\delta(x)$ is fixed once the sample x is given (or there is no randomness conditional on a given x). A more general scenario is the case where an action is not fully fixed when a sample x is given, and instead, action follows a probability distribution $\delta(.|x)$ on $\mathcal{B}_A$. This means that given a sample x, for any set $D \in \mathcal{B}_A$, the probability of taking an action in D is $\delta(D|x)$. As long as the probabilistic mechanism $\delta(\cdot|x)$ is specified, action a is taken according to a randomized decision rule $\delta(\cdot|\cdot)$ that is a function defined on $\mathcal{B}_A \times \mathcal{X}$ and takes values in $[0, 1]$. Obviously, a deterministic decision function is a special case of a randomized decision function. With no exception, a randomized decision function $\delta(\cdot|\cdot)$ is required to satisfy the following condition: For any $D \in \mathcal{B}_A$, $\delta(D|x)$ as a function of x is $\mathcal{B}_x$-measurable.

Randomized decision functions are proposed to provide multiple decisions for consideration with respective favorable scores. This strategy becomes increasingly popular in practice to address uncertainty beyond the available data that prohibits a simple unique decision.

1.2.2 Risk Function

Let us first consider the case of deterministic decision function δ. Let L be the loss function. When sample x is available, we take an action $\delta(x)$. Let the true parameter value be θ. Then, the loss is $L(\theta, \delta(x))$. Since sample X follows a distribution P_θ, the mean loss function

is given by

$$R(\theta, \delta) = E_\theta\{L(\theta, \delta(X))\} = \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x). \tag{1}$$

Here $R(\theta, \delta)$ is termed as the *risk function* of δ. It measures the average loss of a decision function δ when the true parameter is θ. The integral in (1) is valid because both loss function L and decision function δ are measurable functions. In the case of randomized decision function, the risk function can be similarly defined as a mean loss. This definition involves a double averaging. The first is to calculate the average loss conditional on a given sample x, namely, $\int_A L(\theta, a)\delta(da|x)$. Then the second average is taken for x under the probability distribution P_θ. This leads to

$$R(\theta, \delta) = \int_{\mathcal{X}} \left\{ \int_A L(\theta, a)\delta(da|x) \right\} dP_\theta(x). \tag{2}$$

Again, the integral in (2) is valid because both loss function L and decision function δ are measurable functions, as discussed above. In the literature of statistical decision theory, the risk function is the most critical quantity in a decision function. This is because the currently popular optimality criteria for the evaluation of decision function are all based on the risk function.

Let us consider a couple of examples for the calculation of risk function.

Example 6 In Example 1, suppose we like to estimate the variance parameter σ^2. Now the action space $A = \{a : a \geq 0\}$. Consider a squared error loss function $L(\theta, a) = (\sigma^2 - a^2)^2$, $\theta = (b, \sigma)$. We adopt the sample variance estimator as the decision function: $\delta(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}$. Since $(n-1)\delta^2(X)/\sigma^2 \sim \chi^2_{n-1}$, the risk function is given by

$$R(\theta, \delta) = E_\theta(\sigma^2 - \delta^2(X))^2 = 2(\sigma^2)^2/(n-1).$$

Instead, if we use a different decision function $\delta_1(x) = \sqrt{\frac{1}{n+1}\sum_{i=1}^n (x_i - \bar{x})^2}$, then the risk function is

$$R(\theta, \delta_1) = E_\theta(\sigma^2 - \delta_1^2(X))^2 = 2(\sigma^2)^2/(n+1).$$

It is interesting to notice that for any $\theta = (b, \sigma)$, we have $R(\theta, \delta_1) < R(\theta, \delta)$.

Example 7 In Example 1, we consider an interval estimation of parameter b. The loss function of an action $a = [a_1, a_2]$ is

$$L(\theta, a) = \{1 - I_{[a_1, a_2]}(b)\} + m(a_2 - a_1),\ \theta = (b, \sigma)$$

where $m > 0$ is a constant. Here we adopt the commonly used t-statistic-based interval estimation as a decision function:

$$\delta(x) = \left[\bar{x} - \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2), \bar{x} + \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2)\right] \stackrel{def}{=} [a_1, a_2],$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, and $\alpha \in (0,1)$. We have the following two facts:

$$\begin{aligned} E_\theta\{I_{[a_1(X),a_2(X)]}(b)\} &= P_{b,\sigma}\left\{\left|\frac{\sqrt{n}(\bar{X}-b)}{s}\right| \le t_{n-1}(\alpha/2)\right\} = 1-\alpha; \\ E_\theta(s) &= \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\sigma. \end{aligned}$$

It is easy to obtain that

$$R(\theta, \delta) = \frac{2\sqrt{2}m}{\sqrt{n(n-1)}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\sigma t_{n-1}(\alpha/2) + \alpha.$$

1.3 Criterion for the Evaluation of Decision Function

For a given problem of statistical decision, there are many possible decision functions as candidates. Theoretically speaking, any measurable transformation from $(\mathcal{X}, \mathcal{B}_x)$ to $(A, \mathcal{B}_A)$ is a valid candidate. For instance, in Example 1, to estimate the parameter b, we may use sample mean, sample median or others. Apparently, among the available candidates, we aim to choose the best one, or to choose one that is as close to the best as possible. The question is how to determine or evaluate a decision function to be better than others? Answering to this question relies on a certain criterion of optimality for evaluation. There are several criteria proposed in the literature, which are discussed in turn in this section.

1.3.1 Uniform Superiority

A decision rule or decision function δ^* is said to be *uniformly superior or equivalent* to another decision rule δ if

$$R(\theta, \delta^*) \le R(\theta, \delta), \ \forall \theta \in \Theta. \tag{3}$$

If the inequality in (3) holds strictly for at least one θ value, then δ^* is *uniformly superior* to δ. Sometimes, we say that δ^* (uniformly) *dominates* δ. If a decision rule or decision function δ^* is uniformly superior or equivalent to any other decision rules δ's given in a specific statistical problem, then δ^* is said to be *the uniformly optimal decision function.* In the case

where the risk function is the instrument to evaluate the performance of decision function, the uniformly optimal decision function would be the "dream" solution. Unfortunately, such ideal optimal solution only exists in a very handful of cases, and thus this criterion has no practical value.

Example 8 Let $X_1, \ldots, X_n$ be *i.i.d.* with $N(\mu, \sigma^2), \mu \in (-\infty, \infty), \sigma > 0$. The objective is to obtain an esimator of μ under the squared error loss function $L((\mu, \sigma^2), a) = (\mu - a)^2$. In the following we like to show that there does not exist a uniformlly superior estimator of μ.

Let $\delta^*(X)$ be a uniformly superior estimator of μ. Its risk is given by

$$\begin{aligned} R(\mu, \delta^*(X)) &= E_\mu \left[(\mu - \delta^*(X))^2 \right] \\ &= E_{\mu|} \left[(\mu - E(\delta^*(X)) + E_\mu(\delta^*(X)) - \delta^*(X))^2 \right] \\ &= (\mu - E_\mu(\delta^*(X)))^2 + Var_\mu(\delta^*(X)) \end{aligned}$$

Then $R(\mu, \delta^*(X)) \leq R(\mu, \delta(X))$ for all $\mu \in \mathbb{R}$ for any decision rule $\delta \in A$.

Now take a desion rule $\delta(X) = c$ with an arbitrary known constant c, whose risk function is $(\mu - c)^2$. Thus, at $\mu = c$, $R(c, \delta^*(X)) = 0$. This implies that

$$\begin{aligned} 0 &= (c - E_c(\delta^*(X)))^2 + Var_c(\delta^*(X)) \\ \Rightarrow Var_c(\delta^*(X)) &= 0, \quad E_c(\delta^*(X)) = c \\ \Rightarrow \delta^*(X) &= c \end{aligned}$$

Thus there does not exist one instance at which the risk function of δ^* is strickly smaller than that of $\delta = c$. Therefore there cannot be a uniformly superior estimator $\delta^(X)$ of μ.

To overcome this issue, there are typically two strategies suggested in the literature. One is to reduce and/or restrict the scope of decision functions to be considered by adding some eligibility conditions on decision functions to be qualified as candidates; the other is to relax the optimality definition from the tough point-to-point comparison to a less stringent form (e.g. the minimax principle).

1.3.2 Admissibility, Complete Classes and Essentially Complete Classes

The concept of admissibility is proposed to carry out the so-called "reduction principle"; that is, to reduce the scope of decision functions by identifying and eliminating those that

are definitely bad or inadmissible.

Admissibility Let δ be a decision function. If there exists a decision function δ' that dominates δ, then δ is said to be *inadmissible*. On contrary, if there does not exist any decision function that dominates δ, then decision function δ is said to be *admissible*. This concept is intuitively sensible: if there is a uniformly superior decision rule δ' over δ, there is no basis to include δ in the set of candidates. Of course, this exclusion criterion is based on the risk function. In general, it is not easy to determine if a decision function is admissible or inadmissible, and there are some results in the literature of point estimation.

Example 9 Consider Example 6. It is clear that the decision function $\delta_1(x)$ has a strictly smaller risk function than that of the decision function based on the sample variance, $\delta(x)$, for all parameters $(b, \sigma) \in \Theta$. This implies that δ_1 dominates δ, and thus, $\delta(x)$ is inadmissible.

Example 10 Consider Example 8. Let show that estimator $\hat{\mu}(\mathbf{x}) = c\bar{x}$ with $|c| > 1$ is not admissible, where $\bar{x}$ is the sample mean. Let $\delta(X) = c\bar{X}$ with $|c| > 1$. We calculate the risk of $\delta(X)$:

$$\begin{aligned} E_\mu\left[(\mu - \delta(X))^2\right] &= E_\mu\left[\left(\mu - c\bar{X}\right)^2\right] \\ &= E_\mu\left[\left(\mu - c\mu + c\mu - c\bar{X}\right)^2\right] \\ &= (\mu - c\mu)^2 + c^2 Var_\mu\left(\bar{X}\right) \\ &= \mu^2(1-c)^2 + c^2\sigma^2/n \\ &> \sigma^2/n = R(\mu, \bar{X}), \end{aligned}$$

since $|c| > 1$. That is, $\bar{X}$ dominates the $\delta(X)$. Hence $\delta(X) = c\bar{X}$ is not admissible.

Strikingly, we will revisit this point later after the empirical Bayes is intruduced, where we will show that the James-Stein estimator dominates the MLE under the squared error loss.

Complete Classes Let $\mathcal{D}$ be a class of decision functions. For example, it may be a class of all valid decision functions, a class of all valid deterministic decision functions, or a class of all valid decision functions that can be expressed as a linear function of sample. Let

$\mathcal{D}^*$ is a subclass of $\mathcal{D}$. If for any decision function $\delta \in \mathcal{D}\backslash\mathcal{D}^*$ (or $\mathcal{D}^{*c}$), there must exist a $\delta^* \in \mathcal{D}^*$ such that δ^* dominates δ, then $\mathcal{D}^*$ is said to be *complete* with respect to $\mathcal{D}$. In this case, $\mathcal{D}$ can be reduced to a smaller class $\mathcal{D}^*$, because any decision function outside $\mathcal{D}^*$ is inadmissible.

Essentially Complete Classes This is a less stringent concept based on $\leq$ (superior or equivalent to) in the comparison of the risk function. Let $\mathcal{D}$ be a class of decision functions. Let $\mathcal{D}^*$ is a subclass of $\mathcal{D}$. If for any decision function $\delta \in \mathcal{D}\backslash\mathcal{D}^*$ (or $\mathcal{D}^{*c}$), there must exist a $\delta^* \in \mathcal{D}^*$ such that δ^* is superior or equivalent to δ, then $\mathcal{D}^*$ is said to be *essentially complete* with respect to $\mathcal{D}$. In this case, it is loss free to reduce $\mathcal{D}$ to a smaller class $\mathcal{D}^*$. From a theoretical point of view, it is easier to prove a class to be essentially complete than to be complete.

1.3.3 The Minimax Principle

Let δ be a decision function with the risk function equal to $R(\theta, \delta)$. Then, the maximum risk with respect to the use of the decision function δ is given by

$$M(\delta) = \sup_{\theta \in \Theta} R(\theta, \delta).$$

This represents the worst that can happen when rule δ is used. With the desire to protect against the worst possible state of nature, we may consider the following principle: If $M(\delta') < M(\delta)$, then δ' is said to be preferred to δ. Consider a class of decision functions, denoted by $\mathcal{D}$. A decision rule $\delta^* \in \mathcal{D}$ is said to be a minimax solution of a statistical decision problem if $M(\delta^*) \leq M(\delta), \forall \delta \in \mathcal{D}$. In other words, under the minimax principle, there exists no decision function that is better than δ^*.

In the context of parameter estimation, such decision function δ^* is also called the minimax estimator. It is interesting to notice that the minimax principle is NOT based on a point-to-point risk comparison over $\theta \in \Theta$. Rather, it bears on a global value to evaluate the performance of decision function.

1.3.4 The Bayes Decision Principle

Similar to the minimax principle, the Bayes decision principle avoids point-to-point risk comparison and intends to evaluate the performance of decision function using a global

value. This global value used is indeed the mean. This proceeds as follows. Assume there is a probability measure $d\Pi(\theta)$ (i.e. a prior distribution) on a measurable space $(\Theta, \mathcal{B}_\Theta)$. Then we calculate the expectation with respect to the probability measure as follows:

$$R_\Pi(\delta) = \int_\Theta R(\theta, \delta) d\Pi(\theta).$$

This may be regarded as a weighted average of the risk function R according to the probability measure $d\Pi$. Consider two decision functions δ and δ'. A decision rule δ' is said to be preferred to a decision rule δ (in terms of the Bayes decision principle with prior $d\Pi$), if $R_\Pi(\delta') < R_\Pi(\delta)$. Consider a class of decision functions $\mathcal{D}$. A decision rule $\delta_\Pi \in \mathcal{D}$ is said to be a Bayes solution with the prior distribution $d\Pi$, if $R_\Pi(\delta_\Pi) \leq R_\Pi(\delta), \forall \delta \in \mathcal{D}$. In the use of the Bayes decision principle, the choice of the prior distribution $d\Pi(\theta)$ pertains to some subjectivity, which has attracted lot of attentions from researchers in the literature of Bayes Statistics.

1.3.5 Unbiasedness Principle and Invariance Principle

Both principles are intended to reduce the scope of decision functions by imposing certain properties on them.

Unbiasedness Principle See Example 6, where $\delta_1(x)$ is not unbiased. Unbiasedness is a familiar concept. Suppose the statistical problem is to estimate $g(\theta)$ (g being known). We plan to use $\delta(x)$ as an estimator. If for every distribution P_θ in the distribution family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, we have

$$E_\theta\{\delta(X)\} = \int_\mathcal{X} \delta(x) dP_\theta(x) = g(\theta),$$

then δ is an unbiased estimator of $g(\theta)$. If we only consider the class of unbiased estimators (or decision functions), then the range of decision functions reduces greatly, so some theoretical results such as UMVUE may be relatively easily establised. A more critical issue is, perhaps, whether such reduction would lead to a lost opportunity of identifying better decision functions. Although the popularity of unbiasedness in the classical statistical theory, in recent years there is a retreat on its dominance, in particular from the point of view of statistical decision theory.

Invariance Principle Invariance Principle and Equivalence Principle are used exchangeably in the literature. This principle, similar to the unbiasedness principle, aims to reduce the scope of decision functions. Let us begin with a simple example before formally introduce this concept.

Consider the estimation of mean parameter b in Example 1. Let the collected sample be $x_1, \ldots, x_n$. If we change the origin (0) of the measure by unit $-c$, then the sample becomes $x_1 + c, \ldots, x_n + c$, and the resulting mean parameter then becomes $b + c$. Note that the statistical problem remains the same: estimation of b. In the original measure system (or coordination system), an estimator of b is $\delta(x_1, \ldots, x_n)$; in the adjusted coordination system, $\delta(x_1+c, \ldots, x_n+c)$ estimates $b+c$. If we want the estimator remains invariant to the change of origin, then we require

$$\delta(x_1 + c, \ldots, x_n + c) = \delta(x_1, \ldots, x_n) + c, \ \forall c.$$

Obviously, not every estimator satisfies the above property. Thus, imposing this property of invariance on estimators can help us significantly reduce the scope of decision functions, so it is more likely to determine an optimal estimator (decision function) within a smaller class of estimators (decision functions).

Renewal Principle To address the velocity of data arrival in the modern big data collection, in particular the so-called streaming data or tracking data, recently we proposed a new property for decision rule, namely *the renewal principle*. Suppose streaming data are collected in a sequential fashion over daily, weekly or monthly waves. It is desirable if a desicion rule $\delta(\cdot)$ may be renewed or updated with a new arrival data without using individual data points of all past data except the available actions or decisions made with the past data. Suppose $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n})$ is the old data batch and $\mathbf{x}_2 = (x_{21}, \ldots, x_{2m})$ is the new data batch. When we use the sample mean as a decision rule, $\delta(\mathbf{x}_1) = \sum_{i=1}^{n} x_{1i}/n$ is the action based on the old data batch, and after this calculation this old data is no longer accessible in the subsequent computations. This decision rule allows us to renew our action easily based on only the previous action $\delta(\mathbf{x}_1)$ and the new data batch $\mathbf{x}_2$: First sample mean $\delta(\mathbf{x}_2)$ is calculated with the new data batch, and then the decision rule is renewed with both $\mathbf{x}_1$ and

$\mathbf{x}_2$ by the following formula:

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \{n\delta(\mathbf{x}_1) + m\delta(\mathbf{x}_2)\}/(n+m).$$

However, if we choose sample median as the decision rule, then it does not have such a renewal property as median is a calculated according to order statistics that have to involve all individual observations from both the old and the new samples. We recently showed that the MLE in the generalized linear models is a renewable decision rule.

Ethical Principle More and more machine-based automatic systems or apps are available to make decisions on the behalf of us. Are those decision rules derived from machine learning algorithms ethical? For example, if is a credit score given by a machine algorithm fair in the sense that, say, there is no bias to age?