5 Decision Theory: Basic Concepts

Point estimation of an unknown parameter is generally considered the most basic inference problem. Speaking generically, if θ is some unknown parameter taking values in a suitable parameter space Θ, then a point estimate is an educated guess at the true value of θ. Now, of course, we do not just guess the true value of θ without some information. Information comes from data; some information may also come from expert opinion separate from data. In any case, a point estimate is a function of the available sample data. To start with, we allow any function of the data as a possible point estimate. Theory of inference is used to separate the good estimates from the not so good or bad estimates. As usual, we will start with an example to help us understand the general definitions.

Example 5.1. Suppose $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda), \lambda > 0$. Suppose we want to estimate the parameter λ. Now, of course, $\lambda = E_\lambda(X_1)$, i.e., λ is the population mean. So, just instinctively, we may want to estimate λ by the sample mean $\bar{X} = \frac{X_1+\cdots+X_n}{n}$; and, indeed, $\bar{X}$ is a *possible* point estimator of λ. While λ takes values in $\Theta = (0, \infty)$, the estimator $\bar{X}$ takes values in $\mathcal{A} = [0, \infty)$; $\bar{X}$ can be equal to zero! So, the set of possible values of the parameter and the set of possible values of an estimator need not be identical. We must allow them to be different sets, in general.

Now, $\bar{X}$ is certainly not the only possible estimator of λ. We can use essentially any function of the sample observations $X_1, \cdots, X_n$ to estimate the parameter λ. For example, just X_1, or $X_1 + X_2 - X_3$, or even seemingly poor estimators like X_1^4. Any estimator is allowed to begin with; theory will separate the good ones from the bad ones.

Next, suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ, σ are unknown parameters. So, now, we have got a two dimensional parameter vector, $\theta = (\mu, \sigma)$. Suppose we want to estimate μ. Once again, a possible estimator is $\bar{X}$; a few other possible estimators are the sample median, $M_n = \text{median}\{X_1, \cdots, X_n\}$, or $\frac{X-2+X_4}{2}$, or even seemingly poor estimators, like $100\bar{X}$.

Suppose, it was *known to us* that μ must be nonnegative. Then, the set of possible values of μ is $\Theta = [0, \infty)$. However, the instinctive point estimator $\bar{X}$ *can take any real value.* It takes values in the set $\mathcal{A} = (-\infty, \infty)$. You would notice again that $\mathcal{A}$ is not the same as Θ in this case. In general, $\mathcal{A}$ and Θ can be different sets.

If we want instead to estimate σ^2, which is the population variance, a first thought is to use the *sample variance* $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ (dividing by $n-1$ rather than n seems a little odd at first glance, but has a mathematical reason, which will be clear soon). In this example, the parameter σ^2 and the estimator s^2 both take values in $(0, \infty)$. But, if we knew that $\sigma^2 \leq 100$, say, then once again, Θ and $\mathcal{A}$ would be different. And, as always, there are many other possible estimators of σ^2, for example, $\frac{1}{n-1}\sum_{i=1}^n (X_i - M_n)^2$, where M_n is the sample median. Here is a formal definition of a point estimator.

Definition 5.1. Let the vector of sample observations $X^{(n)} = (X_1, \cdots, X_n)$ have a joint distribution $P = P_n$ and let $\theta = h(P)$, taking values in the parameter space $\Theta \subseteq \mathcal{R}^p$, be a parameter of the distribution P. Let $T(X_1, \cdots, X_n)$ taking values in a specified set $\mathcal{A} \subseteq \mathcal{R}^p$ be a general statistic, Then, any such $T(X_1, \cdots, X_n)$ is called a *point estimator* of θ.
The set Θ is called *the parameter space*, and the set $\mathcal{A}$ is called *the statistician's action space.*
If specific observed sample data $X_1 = x_1, \cdots, X_n = x_n$ are available, then the particular value $T(x_1, \cdots, x_n)$ is called an *estimate* of θ. *Thus, the word estimator applies to the general function $T(X_1, \cdots, X_n)$, and the word estimate applies to the value $T(x_1, \cdots, x_n)$ for specific data. In this text, we use estimator and estimate synonymously.*
A standard general notation for a generic estimate of a parameter θ is $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$.

5.0.1 Evaluating an Estimator and MSE

Except in rare cases, the estimate $\hat{\theta}$ would not be exactly equal to the true value of the unknown parameter θ. It seems reasonable that we like an estimate $\hat{\theta}$ which generally comes quite close to the true value of θ, and dislike an estimate $\hat{\theta}$ which generally misses the true value of θ by a large amount. How, are we going to make this precise and quantifiable?
The general approach to this question involves the specification of a loss and a risk function, which we will introduce in a later section. For now, we describe a very common and even hugely popular criterion for evaluating a point estimator, the mean squared error.

Definition 5.2. Let θ be a real valued parameter, and $\hat{\theta}$ an estimate of θ. The mean squared error (MSE) of $\hat{\theta}$ is defined as

$$MSE = MSE(\theta, \hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2], \ \theta \in \Theta.$$

If the parameter θ is p-dimensional, $\theta = (\theta_1, \cdots, \theta_p)$, and $\hat{\theta} = \hat{\theta}_1, \cdots, \hat{\theta}_p)$, then the mean squared error f $\hat{\theta}$ is defined as

$$MSE = MSE(\theta, \hat{\theta}) = E_\theta[||\hat{\theta} - \theta||^2] = \sum_{i=1}^{p} E_\theta[(\hat{\theta}_i - \theta_i)^2], \ \theta \in \Theta.$$

5.0.2 Bias and Variance

It turns out that the mean squared error of an estimator has one component to do with *systematic error* of the estimator and a second component to do with *random error* of the estimator. If an estimator $\hat{\theta}$ routinely overestimated the parameter θ, then usually we will have $\hat{\theta} - \theta > 0$; we think of this as the estimator making a systematic error. Systematic

errors can also be made by routinely underestimating θ, We quantify the systematic error of an estimator by looking at $E_\theta(\hat{\theta}-\theta) = E_\theta(\hat{\theta})-\theta$; this is called the *bias* of the estimator; we denote it as $b(\theta)$. If the bias of an estimator $\hat{\theta}$ is always zero, $b(\theta) = 0$ for all θ, then the estimator $\hat{\theta}$ is called *unbiased*. On the other hand, the estimator $\hat{\theta}$ may not make much systematic error, but still can just be unreliable because from one dataset to another, its accuracy may differ wildly. This is called *random or fluctuation error*, and we often quantify the random error by looking at the variance of the estimator, $\text{Var}_\theta(\hat{\theta})$. A pleasant property of the MSE of an estimator is that always the MSE neatly breaks into two components, one involving the bias, and the other involving the variance. You have to try to keep both of them small; large biases and large variances are both red signals. Here is a bias-variance decomposition result.

Theorem 5.1. Let θ be a real valued parameter and $\hat{\theta}$ an estimator with a finite variance under all θ. Then,

$$MSE(\theta, \hat{\theta}) = \text{Var}_\theta(\hat{\theta}) + b^2(\theta),\ \theta \in \Theta.$$

Proof; To prove this simple theorem, we recall the elementary probability fact that for any random variable U with a finitte variance, $E(U^2) = \text{Var}(U) + [E(U)]^2$. Identifying U with $\hat{\theta} - \theta$,

$$\begin{aligned} MSE(\theta, \hat{\theta}) &= \text{Var}_\theta(\hat{\theta} - \theta) + [E(\hat{\theta} - \theta)]^2 \\ &= \text{Var}_\theta(\hat{\theta}) + b^2(\theta). \end{aligned}$$

5.0.3 Computing and Graphing MSE

We will now see one introductory example.

Example 5.2. (Estimating a Normal Mean and Variance) Suppose we have sample observations $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where $-\infty < \mu < \infty, \sigma > 0$ are unknoown parameters. The parameter is two dimensional, $\theta = (\mu, \sigma)$.

First consider estimation of μ, and as an example, consider these two estimates: $\bar{X}$ and $\frac{n\bar{X}}{n+1}$. We will calculate the MSE of each estimate, and make some comments.

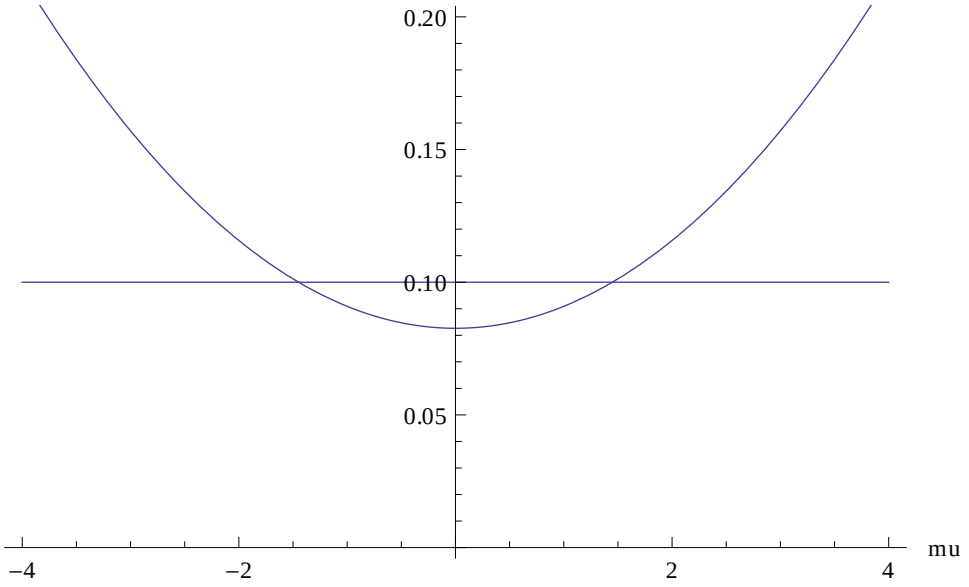
Since $E(\bar{X}) = \frac{1}{n} sum_{i=1}^n E(X_i) = \frac{1}{n}(n\mu) = \mu$ for any μ and σ, the bias of $\bar{X}$ for estimating μ is zero; $E_\theta(\bar{X}) - \mu = \mu - \mu = 0$. In other words, $\bar{X}$ is an unbiased estimate of μ. Therefore, the MSE of $\bar{X}$ is just its variance,

$$E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \frac{\text{Var}(X_1)}{n} = \frac{\sigma^2}{n}.$$

Notice that the MSE of $\bar{X}$ does not depend on μ; it only depends on σ^2.

Now, we will find the MSE of the other estimate $\frac{n\bar{X}}{n+1}$. For this, by our Theorem 3.1, the MSE of $\frac{n\bar{X}}{n+1}$ is

$$E[(\frac{n\bar{X}}{n+1} - \mu)^2] = \text{Var}(\frac{n\bar{X}}{n+1}) + [E(\frac{n\bar{X}}{n+1} - \mu)]^2$$



$$= (\frac{n}{n+1})^2 \text{Var}(\bar{X}) + [\frac{n}{n+1}\mu - \mu]^2 = (\frac{n}{n+1})^2 \frac{\sigma^2}{n} + (\frac{1}{n+1})^2 \mu^2$$
$$= \frac{\mu^2}{(n+1)^2} + \frac{n\sigma^2}{(n+1)^2}.$$

Notice that the MSE of this estimate *does depend on both* μ *and* σ^2; $\frac{\mu^2}{(n+1)^2}$ is the contribution of the bias component in the MSE, and $\frac{n\sigma^2}{(n+1)^2}$ is the contribution of the variance component.

For purposes of comparison, we plot the MSE of both estimates, taking $n = 10$ and $\sigma = 1$; the MSE of $\bar{X}$ is constant in μ, and the MSE of $\frac{n\bar{X}}{n+1}$ is a quadratic in μ. For μ near zero, $\frac{n\bar{X}}{n+1}$ has a smaller MSE, but otherwise, $\bar{X}$ has a smaller MSE. The graphs of the two mean squared errors cross. This is quite often the case in point estimation.

Next, we consider estimation of σ^2. Now, we will use as our example the estimates $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. First, we will prove that s^2, i.e., the estimator that divides by $n-1$ is an unbiased estimator of σ^2. Here is the proof. First, note the algebraic identity

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2) = \sum_{i=1}^n X_i^2 - 2\bar{X}\sum_{i=1}^n X_i + n\bar{X}^2$$
$$= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2.$$

Therefore,

$$E[\sum_{i=1}^n (X_i - \bar{X})^2] = nE(X_1^2) - nE[\bar{X}^2] = n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) = (n-1)\sigma^2,$$

which gives $E(s^2) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$. Therefore, the MSE of s^2, by Theorem 3,1, is the same as its variance,

$$E[(s^2-\sigma^2)^2] = \text{Var}(s^2) = \text{Var}(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2)$$

$$= \frac{1}{(n-1)^2}\text{Var}(\sigma^2\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}) = \frac{\sigma^4}{(n-1)^2}\text{Var}(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2})$$

$$= \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1},$$

by using the fact that if $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$, then $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$ and that the variance of a χ^2 distribution is twice its degrees of freedom.
Next, we will fiind the MSE of the second estimate $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$. This estimate does have a bias! Its bias is

$$b(\theta) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Also, its variance is

$$\text{Var}(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2) = \text{Var}(\frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2)$$

$$= \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2\sigma^4(n-1)}{n^2}.$$

Therefore, by Theorem 3.1, the MSE of our second estimate $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ is

$$E[(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 - \sigma^2)^2] = \frac{2\sigma^4(n-1)}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

You may verify that this is always smaller than $\frac{2\sigma^4}{n-1}$, the MSE of s^2. So, we have the scenario that in this example, the *biased estimator* $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ always has a smaller MSE than the *unbiased estimator* s^2. This is a special example; do not expect this to automatically hold in another example.

5.0.4 Loss and Risk Functions

Use of MSE is a special case of a more general formulation of estimation as a *decision theory problem*. In this formulation, we view the point estimation problem as a two person game, player I who chooses the true value of θ from Θ, and player II who chooses a point estimate, possibly after observing some sample data. Player I is called *nature*, and player II *the statistician*. In accordance with our previous notation, we allow the statistician to choose the value of his estimate from a well defined set $\mathcal{A}$, called the *action space*. There is

another component necessary to complete the formulation. After player I chooses θ from Θ, and player II chooses a specific estimate for θ, say $\hat{\theta}(x_1, \cdots, x_n) = a$, player II has to pay player I a *penalty* for making a possibly incorrect guess on the true value of θ. This is called *the loss function*, and is denoted as $L(\theta, a), \theta \in \Theta, a \in \mathcal{A}$. In principle, the loss function can bbe almost any function with some natural constraints; but, in our practice of statistics, we tend to use just one or two most of the times, because it is easier to work with them.

We note that the *realized value* of an estimate $\hat{\theta}(x_1, \cdots, x_n)$ depends on the particular data $x_1, \cdots, x_n$ actually obtained. In one sampling experiment, our data may be such that the estimate $\hat{\theta}(x_1, \cdots, x_n)$ comes quite close to the true value of θ, while in another independent sampling experiment, the data may be such that the estimate is not so good. We try to look at average performance of $\hat{\theta}(X_1, \cdots, X_n)$ *as an estimating procedure, or decision rule*. This motivates the definition of a *risk*. One comment is now in order. Since many other inference problems besides point estimation can be put under the general formulation of decision theory, we prefer to use a single notation for the statistician's decision procedure. To be consistent with much of the statistical literature, we use the notation $\delta(X_1, \cdots, X_n)$. In the point estimation context, $\delta(X_1, \cdots, X_n)$ will mean an estimator; in another context, it may mean something else, like a *test*.

Definition 5.3. Let Θ denote nature's parameter space, $\mathcal{A}$ the statistician's action space, and $L(\theta, a)$ a specific loss function. Then, the *risk function* of a decision procedure $\delta(X_1, \cdots, X_n)$ is the average loss incurred by δ:

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X_1, \cdots, X_n))].$$

In the above, the expectation E_θ means expectation with respect to the joint distribution of $(X_1, \cdots, X_n)$ under θ.

Example 5.3. (Some Possible Loss Functions) We usually impose the condition that $L(\theta, a) = 0$ if $a = \theta$ (no penalty for perfect work). We also often take $L(\theta, a)$ to be a monotone nondecreasing function of the physical distance between θ and a. So, if θ and a are real valued, we often let $L(\theta, a) = W(|theta - a|)$ for some function W with $W(0) = 0$, and $W(x)$ monotone nondecreasing on $[0, \infty)$.

Some common choices are:

$$\underline{\text{Squared error loss}} \quad L(\theta, a) = (a - \theta)^2;$$

$$\underline{\text{Absolute error loss}} \quad L(\theta, a) = |a - \theta|;$$

$$\underline{\text{Zero-K loss}} \quad L(\theta, a) = KI_{|a-\theta|>c};$$

$$\underline{\text{Asymmetric loss}} \quad L(\theta, a) = K_1 \text{ if } a < \theta, = 0 \text{ if } a = \theta, K_2 \text{ if } a > \theta;$$

$$\underline{\text{Power loss}} \quad L(\theta, a) = |a - \theta|^\alpha, \alpha > 0;$$

$$\underline{\text{Weighted squared error loss}} \quad L(\theta, a) = w(\theta)(a - \theta)^2.$$

In a practical problem, writing down a loss function that correctly reflects consequences of various decisions under every possible circumstance is difficult and even impossible to do. You should treat decision theory as a frequently useful general guide as to the choice of your procedure. By far, in point estimation, squared error loss is the most common, because it is the easiest to work with, and gives reasonable general guide as to which procedures are good and which are bad. Note the important fact that

Risk function under squared error loss is the same as MSE.

5.0.5 Optimality and Principle of Low Risk

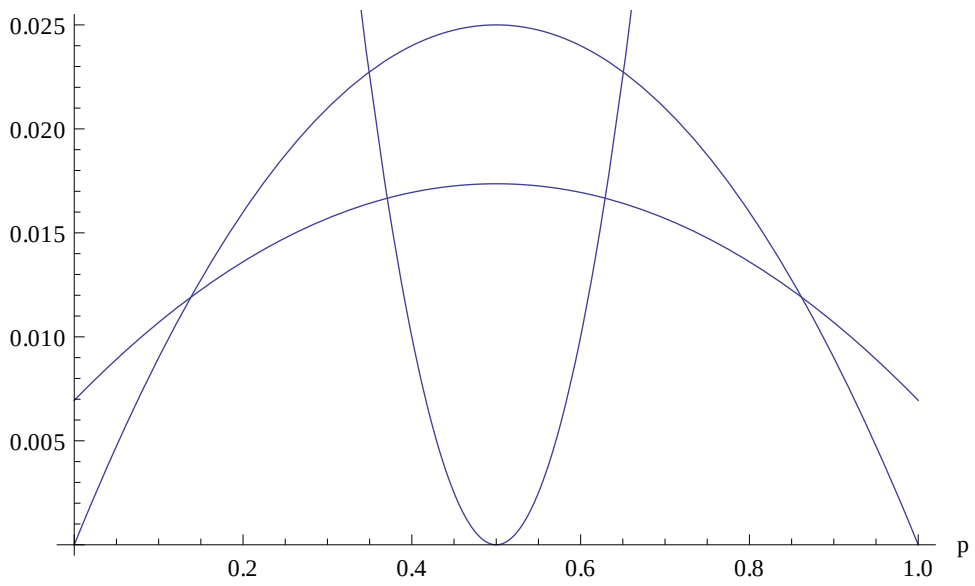
It seems only natural that once we have specified a loss function, we should prefer decision procedures δ that have low risk. Lower risk is considered such a good property, that procedures whose risk functions can always be beaten by some other alternative procedure are called *inadmissible*. In fact, it would be the best if we could find one decision procedure δ_0 which has *the smallest risk among all possible decision procedures at every possible value of* θ. Unfortunately, rarely, this is possible. There are no uniformly best decision procedures. Risk functions of decent decision procedures tend to cross; sometimes, one is better and some other times, the other procedure is better. In the next few sections, we will give an elementary introduction to popular methods for choosing among decision procedures whose risk functions cross. But, right now, let us see one more concrete risk calculation.

Example 5.4. (Estimation of Binomial p). Suppose n observations $X_1, \cdots, X_n$ are taken from a Bernoulli distribution with an unknown parameter $p, 0 < p < 1$. Suppose the statistician's action space is the closed interval $[0, 1]$, and that we use the squared error loss function $(a - p)^2$. Let $T = \sum_{i=1}^n X_i$ be the total number of successes. We consider three decision procedures (estimators)

$$\delta_1(X_1, \cdots, X_n) = \frac{T}{n};$$

$$\delta_2(X_1, \cdots, X_) = \frac{T + 1}{n + 2};$$

$$\delta_3(X_1, \cdots, X_n) \equiv \frac{1}{2}.$$



Note that all three estimators are of the form $a+bT$ for suitable a, b. The bias of a general estimator of this form is $b(p) = a + bnp - p = a + (nb-1)p$, and so, the risk function of a general estimator of this form is

$$R(p, a+bT) = E[(a+bT-p)^2] = b^2(p) + \mathrm{Var}(a+bT)$$

$$= (a+(nb-1)p)^2 + nb^2p(1-p) = a^2 + (nb^2+2abn-2a)p + (n^2b^2 - nb^2 - 2nb+1)p^2. \quad (1)$$

δ_1 corresponds to $a = 0, b = \frac{1}{n}$, δ_2 to $a = \frac{1}{n+2}, b = \frac{1}{n+2}$, and δ_3 to $a = \frac{1}{2}, b = 0$. Substitution into (3) yields:

$$R(p, \delta_1) = \frac{p(1-p)}{n};$$

$$R(p, \delta_2) = \frac{1+(n-4)p(1-p)}{(n+2)^2};$$

$$R(p, \delta_3) = \frac{1}{4} - p(1-p).$$

These three risk functions are plotted above. Evidently, none of them is uniformly the best among the three estimators. Near $p = 0, 1, \delta_1$ is the best, near $p = \frac{1}{2}, \delta_3$ is the best, and in two subintervals away from $0, 1$ and away from $\frac{1}{2}$, δ_2 is the best. We must somehow collapse each of these risk functions into a single representative number, so that we can compare numbers rather than functions. We present some traditional ideas on how to collapse risk functions to a single representative number in our next two sections.

5.0.6 Prior Distributions and Bayes Procedures: First Examples

An intuitively attractive method to collapse a risk function into a single number is to take an average of the risk function with respect to a weight function, say $\pi(\theta)$. For example, in our above binomial example, the parameter is p, and we average the risk function of any procedure δ by calculating $\int_0^1 R(p,\delta)\pi(p)dp$, for a specific weight function $\pi(p)$. Usually, we choose $\pi(p)$ to be a probability density function on $[0,1]$, the parameter space for p. Such weight functions are called *prior distributions* for the parameter, and the average risk $r(\pi,\delta) = \int_0^1 R(p,\delta)\pi(p)dp$ is called the *Bayes risk* of the procedure δ with respect to the prior π. The idea is to prefer a procedure δ_2 to δ_1 if $r(\pi,\delta_2) < r(\pi,\delta_1)$. Since we are no longer comparing two functions, but only comparing two numbers, it is usually the case that one of the two numbers will be strictly smaller than the other one, and that will tell us which procedure is to be preferred, *provided that we are sold on the choice of the specific prior distribution* π that we did choose. This approach is known as *Bayesian decision theory* and is a very important component of statistical inference. The Bayes approach will often give you useful ideas about which procedures may be reasonable in a given problem.
We now give the formal definitions.

Definition 5.4. Let G be a specific probability distribution on natures's parameter space Θ; G is called a *prior distribution* or simply a prior for the parameter θ.

Definition 5.5. Let $R(\theta,\delta)$ denote the risk function of a general decision procedure δ under a specific loss function $L(\theta,a)$. The *Bayes risk* of δ with respect to the prior G is defined as

$$r(G,\delta) = E_G[R(\theta,\delta)],$$

where the notation E_G means an expectation by treating the parameter θ as a random variable with the distribution G. If θ is a continuous parameter and the prior G has a density function π, we have

$$r(\pi,\delta) = \int_\Theta R(\theta,\delta)\pi(\theta)d\theta.$$

Remark: *Although we do not look kindly upon such procedures δ, the Bayes risk $r(G,\delta)$ can be infinite.*
Since lower risks are preferred in general, we would also prefer lower Bayes risk. This raises an immediate natural question; *which particular δ makes $r(G,\delta)$ the smallest possible among all decision procedures*? This motivates an important definition.

Definition 5.6. Given a loss function L and a prior distribution G, any procedure δ_G

such that

$$r(G, \delta_G) = \inf_{\delta} r(G, \delta)$$

is called a Bayes procedure under L and G.

Remark: *A Bayes procedure need not always exist, and when it exists it need not in general be unique. But, under commonly used loss functions, a Bayes procedure does exist, and it is even unique. This is detailed later.*

Let us now return to the binomial example to illustrate these various new concepts.

Example 5.5. (Estimation of Binomial p). As an example, take the specific prior distribution $G = U[0,1]$ with $\pi(p) = 1I_{0 \le p \le 1}$. Then, the Bayes risks of the three procedures $\delta_1, \delta_2, \delta_3$ are

$$r(\pi, \delta_1) = \int_0^1 \frac{p(1-p)}{n} dp = \frac{1}{6n};$$

$$r(\pi, \delta_2) = \int_0^1 \frac{1 + (n-4)p(1-p)}{(n+2)^2} dp = \frac{1}{6(n+2)};$$

$$r(\pi, \delta_3) = \int_0^1 [\frac{1}{4} - p(1-p)] dp = \frac{1}{12}.$$

We now see on easy algebra that if $n > 2$, then an ordering among the three procedures $\delta_1, \delta_2, \delta_3$ has emerged:

$$r(\pi, \delta_2) < r(\pi, \delta_1) < r(\pi, \delta_3).$$

Therefore, *among these three procedures*, we should prefer δ_2 over the other two if $n > 2$ and *if we use* $G = U[0,1]$ *as the prior*. Compare this neat conclusion with the difficulty of reaching a conclusion by comparing their risk functions, which cross.

Remark: It may be shown that δ_2 is in fact the unique Bayes procedure in the Binomial p example, if loss is squared error and $G = U[0,1]$. It will not be proved in this chapter. However, if you change the choice of the prior G to some other distribution on $[0,1]$, then the Bayes procedure will no longer be δ_2. It will be some other procedure, depending on exactly what the prior distribution G is; Bayes procedures are to be treated in detail in Chapter 7.

5.0.7 Maximum Risk and Minimaxity

A second common method to collapse a risk function $R(\theta, \delta)$ to a single number is to look at the maximum value of the risk function over all possible values of the parameter θ. Precisely, consider the number

$$\bar{R}(\delta) = \max_{\theta \in \Theta} R(\theta, \delta).$$

In this approach, you will prefer δ_2 to δ_1 if $\bar{R}(\delta_2) < \bar{R}(\delta_1)$. According to the *principle of minimaxity*, we should ultimately use that procedure δ_0 that results in the minimum

possible value of $\bar{R}(\delta)$ over all possible procedures δ. That is, you minimize over δ the maximum value of $R(\theta, \delta)$; hence the name *minimax*.

Consider a layman's example to understand the concept of minimaxity. Suppose you have to travel from a city A to a small town B which is about 500 miles away. You can either fly or drive. However, only single engine small planes fly to town B, and you are worried about their safety. On the other hand, driving to B will take you 10 hours, and it will cost you valuable time, may be some income, opportunities for doing other pleasurable things, and it would be rather tiring. But you do not think that driving to B can cause you death. The worst possible risk of driving to B is lower than the worst possible risk of flying to B, and according to the minimaxity principle, you should drive.

If we present minimaxity in such a light, it comes across as the philosophy of the timid. But, in many problems of statistical inference, the principle of minimaxity has ultimately resulted in a procedure that is reasonable, or that you would have probably used anyway because of other reasons. It can also serve as an idealistic benchmark against which you can evaluate other procedures. Moreover, although it is not obvious, the approach of minimaxity and the Bayes approach are sometimes connected. A minimax procedure is sometimes a Bayes procedure with respect to a suitable prior G, and this sort of a connection gives some additional credibility to minimax procedures. Brown (1994, 2000) are two very lucidly written articles that explain the role of minimaxity in the evolution of statistical inference.

Let us now revisit our binomial p example for illustration.

Example 5.6. (Estimation of Binomial p). The plot of the risk functions of the three procedures $\delta_1, \delta_2, \delta_3$ reveals what their maximum risks are. The maximum risk of δ_1 and δ_2 is attained at $p = \frac{1}{2}$, and the maximum risk of δ_3 is at the boundary points $p = 0, 1$ (strictly speaking, the maximum is a *supremum*, and the supremum is not attained at any p in the open interval $(0, 1)$). Precisely,

$$\bar{R}(\delta_1) = R(\frac{1}{2}, \delta_1) = \frac{1}{4n};$$

$$\bar{R}(\delta_2) = R(\frac{1}{2}, \delta_2) = \frac{n}{4(n+2)^2};$$

$$\bar{R}(\delta_3) = \lim_{p \to 0,1} R(p, \delta_3) = \frac{1}{4}.$$

We have the ordering

$$\bar{R}(\delta_2) < \bar{R}(\delta_1) < \bar{R}(\delta_3),$$

and therefore, among the three procedures $\delta_1, \delta_2, \delta_3$, according to the principle of minimizing the maximum risk, we should prefer δ_2.

Remark: It may be shown that the overall minimax procedure among all possible procedures is a rather obscure one; it is *not* δ_2. You will see it in Chapter 7. Note that the principle of minimaxity in its formulation does not call for specification of a prior. Some statisticians consider this a plus. However, as we remarked earlier, at the end, a minimax procedure may turn out to be a Bayes procedure with respect to some prior. Very often, this prior is difficult or impossible to guess. Minimax procedures can be extremely difficult to find. It is also important to remember that although minimaxity does not involve the assumption of a specific prior, *absolutely does require the assumption of a specific loss function.* Just as an example, if in our binomial p problem, we use absolute error loss $|p-a|$ instead of squared error loss $(p=a)^2$, the minimax procedure will change, and even worse, no one has ever worked it out!

5.0.8 Assumptions Matter

Statistical inference always requires some assumptions. You have to assume something. How much we assume depends on the problem, on the approach, and perhaps on some other things, like how much data do we have to test our assumptions. For example, for finding a minimax procedure, we have to make assumptions about the model, meaning the distribution of the underlying random variable (normal or Cauchy, for instance), and we have to make an assumption of a specific loss function. In the Bayes approach, too, we will need to make both oof these assumptions and an extra assumption of a specific prior G.

It is important that we at least know what assumptions we have made, and when possible, make some attempt to validate those assumptions by using the available data. Model validation is not an easy task, especially when we have made assumptions which are hard to verify without a lot of data.

Statisticians do not agree on whether prior distributions on parameters can or even should be validated from one's data. There are very few formal and widely accepted methods for verifying a chosen loss function. The original idea was to *elicit* a problem specific loss function by having a personal meeting with the client. Such efforts have generally been regarded as impractical, or have not resulted in clear success. In addition, if loss functions are always problem specific, then we eliminate the premise of a structured theory that will apply simultaneously to many problems. So, when it comes to the assumption of a loss function, we have generally sided with convenience; it is the primary reason that squared error loss and MSE are so dear to us as a profession. We recommend Dawid (1982), Seidenfeld (1985), and Kadane and Wolfson (1998) for thoughtful exposition of these issues.

subsubsectionRobustness as a Theme In statistical inference, the phrase *robust* applies to the somewhat nebulous property of insensitivity of a particular procedure to the assump-

tions that were made. The idea of robust inferences seems to have been first mentioned in those words in Box (1953).

Here is a simple example. You will see in Chapters 6 and 8 that for estimating the location parameter μ of a univariate normal distribution, the sample mean $\bar{X}$ is almost universally regarded as the estimator that one should use. Now, we have made some assumptions here, namely, that the sequence of our sample observations $X_1, \cdots, X_n$ are iid observations from some normal distribution, $N(\mu, \sigma^2)$. uppose now that we did our modelling a little too carelessly. Suppose the observations are iid, but from a distribution with a much heavier tail than the normal; e.g., suppose a much better model would have been that $X_1, \cdots, X_n \overset{iid}{\sim} C(\mu, \sigma)$ for some μ, σ. Then, as it turns out, the sample mean $\bar{X}$ is an extremely poor estimate of μ. You can therefore argue that $\bar{X}$ is *not robust* to the assumption of normality, when it comes to our assumption about the tail. We can similarly ask and investigate if $\bar{X}$ is robust to the assumption of independence of our sample observations $X_1, \cdots, X_n$.

Can we do anything about it? The answer depends on our ambition. *No one can build a procedure which is agreably robust to every assumption that one has made.* However, if we isolate a few assumptions, perhaps only those that we are most unsure about, then it may be possible to construct procedures which are robust against those specific assumptions. For example, if we assume symmetry, but want to be robust against the tail of our distribution, then we can do something about it. As an illustration, if all we wish is to be robust against the tail, then to estimate the location parameter μ of a density on the real line, we would be much better off using the sample median as an estimator of μ, instead of the sample mean. But, if in addition to the tail, we want to be robust against the assumption that our density is symmetric around the parameter μ, then we do not have any reasonable robust solutions to that problem. *We just cannot be simultaneously robust against all our assumptions.* It is also important to understand that robustness is a desirable goal, but not necessarily the first goal. In setting robustness as our first goal, we may end up selecting a procedure which is so focused on caution that it does only a mediocre job in all situations, and not an excellent job in any situation. Robustness is inherently a murky and elusive concept. It is sometimes obtainable to a limited extent. But, still, we must always pay extremely close attention to all of our assumptions.

In doing honest inference, it is important that we know and understand *all the assumptions we have made.* For example, in a point estimation problem, we should ask ourselves:

1. Have we assumed that our sample observations are iid from some F?
2. Have we assumed that F belongs to some specific parametric family, such as Poisson or normal?
3. What is our assumed loss function?

4. What is our assumed prior distribution?

5. Which of these assumptions are seriously questionable?

6. Do we know how to examine if our proposed estimator is reasonably robust to those assumptions?

7. If we conclude that it is not reasonably robust, do we know what to do next?

We will treat robust estimation to some extent in Chapter 7. Some additional references on robust inference are Huber (1981), Hampel et. al (1986), Portnoy and He (2000), and Stigler (2010).

5.1 Exercises

Exercise 5.1. (Skills). Suppose $X_1, \cdots, X_5$ are five iid observations from a Bernoulli distribution with parameter p. Let $T = \sum_{i=1}^{5} X_i$.
(a) Calculate in closed form the MSE of the following three estimators:

$$\delta_1(X_1, \cdots, X_5) = \frac{T}{5};$$

$$\delta_2(X_1, \cdots, X_5) = \frac{T+2}{9};$$

$$\begin{aligned} \delta_3(X_1, \cdots, X_5) &= \tfrac{1}{2}, \qquad \text{if } T = 2, 3 \\ &= \tfrac{T}{5}, \quad \text{if } T = 0, 1, 4, 5. \end{aligned}$$

(b) Plot the MSE of each estimator on a single plot, and comment.

Exercise 5.2. (Partly Conceptual). Suppose X, Y, Z are three iid observations from a normal distribution with mean μ and known variance σ^2. Let U, V, W denote the smallest, the median, and the largest among the three observations X, Y, Z.
(a) Which of the following three estimators of μ are unbiased, i.e., the bias is zero for all μ:

$$V; W; \frac{U+W}{2}; .2U + .6V + .2W.$$

(b) Consider the following strange sounding estimator. Toss a fair coin. If it lands on heads, estimate μ as V; if it lands on tails, estimate μ as $\frac{X+Y+Z}{3}$. Is this estimator unbiased?
(c) Do you know how to calculate the MSE of the estimator in part (b)?

Exercise 5.3. Suppose that we have two observations on an unknown parameter μ. The first one is $X \sim N(\mu, 1)$, and the second one is $Y \sim U[\mu - 1, \mu + 1]$, where X, Y are indepenndent.

(a) Find the MSE of an estimator of μ of the general form $aX + bY$.

(b) When is such an estimator unbiased?

(c) Between the three estimators $X, Y, \frac{X+Y}{2}$, do you have a preference for one of them? Justify your preference.

Exercise 5.4. (Thresholding Estimate). Suppose $X \sim N(\mu, 1)$. Consider the *thresholding estimator* of μ that equals zero if $|X| \le 1$ and equals X if $|X| > 1$.

(a) Find the expectation and the second moment of this estimator.

(b) Hence find the bias and the variance of this estimator.

(c) Plot the MSE of this estimator and the MSE of the estimator X. Do they cross?

(d) Intuitively, when would you prefer to use the thresholding estimate over X?

Exercise 5.5. (Absolute Error Loss). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. If the loss function is the absolute error loss $L(\mu, a) = |\mu - a|$, find the risk function of $\bar{X}$.

Exercise 5.6. (Absolute Error Loss). Suppose a single observation $X \sim Poi(\lambda)$. If the loss function is the absolute error loss $L(\lambda, a) = |\lambda - a|$,

(a) Find the risk of the estimator X when $\lambda = 1$.

(b) Find the risk of the estimator X when $\lambda = 1.4$.

Exercise 5.7. Suppose a single observation $X \sim Poi(\lambda)$. Suggest an estimator of λ^2 and then criticize your own choice.

Exercise 5.8. (Skills). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is considered known. Suppose loss is squared error, and consider estimators of μ of the general form $a + b\bar{X}$.

(a) For what choices of a, b is the risk function uniformly bounded as a function of μ?

(b) For such choices of a, b, find the maximum risk.

(c) Identify the particular choice of a, b that gives the smallest value for the maximum risk. What is the corresponding estimator?

Exercise 5.9. (Cost of Sampling). Suppose a certain Bernoulli parameter is definitely known to be between .46 and .54. It costs some c dollars to take one Bernoulli observation. Would you care to sample, or estimate p to be .5 without bothering to take any data?

Exercise 5.10. (Signal plus Background). Suppose a latent variable $Y \sim Poi(\lambda), \lambda > 0$; Y cannot be directly measured. Instead we measure $X = Y + B$, where $B \sim Poi(2)$, and Y, B are independent. In a physical experiment, your observed value of X turned out to be zero. What would be your estimate of λ?

Exercise 5.11. (Bayes Risks). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$, and suppose λ has a standard exponential prior distribution. Take a squared error loss function.

(a) Calculate the Bayes risk of each of the following estimators of λ:

$$\text{The constant estimate } 1; \bar{X}; \frac{\sum_{i=1}^{n} X_i + 1}{n + 2}.$$

(b) Which of the three estimators has the smallest Bayes risk? The largest Bayes risk?

Exercise 5.12. (Bayes Risks). Suppose $X \sim N(\mu, 1)$ and $Y \sim U[\mu - 1, \mu + 1]$, and that X, Y are independent. Suppose μ has a standard normal prior distribution. Take a squared error loss function.

(a) Calculate the Bayes risk of each of the following estimators of μ:

$$X, Y, \frac{X+Y}{2}.$$

(b) Which estimator has a smaller Bayes risk?

Exercise 5.13. (Bayes Risks). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ^2 is considered known. Suppose loss is squared error. Calculate the Bayes risk of the estimator $\frac{n}{n+1}\bar{X}$ for each of the following prior distributions G:

$$G = U[-3, 3]; N(0, 1); C(0, 1).$$

Exercise 5.14. (Uniform Distributions). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta], \theta > 0$. Take a squared error loss function.

(a) Find the risk function of each of the following three estimators of θ:

$$2\bar{X}; X_{(n)}; \frac{n+1}{n} X_{(n)},$$

where $X_{(n)}$ is the sample maximum.

(b) Calculate the Bayes risk of each of these estimators if θ has a general gamma prior distribution, $\theta \sim G(\alpha, \lambda)$.

(c) Which estimator has the smallest Bayes risk? Does this depend on the values of α, λ?

Exercise 5.15. (Maximum Risk). Suppose $X \sim N(\mu, 1)$, *and it is known that* $-1 \leq \mu \leq 1$. Suppose loss is squared error.

(a) Calculate the risk function of each of the following estimators:

$$X; \frac{X}{2}; \text{the constant estimator zero.}$$

(b) Find the maximum risk of each estimator. Comment.

(c) Consider estimators of the general form $a + bX, -\infty < a < \infty, b \geq 0$. For which choice of a, b is the maximum risk minimized?

Exercise 5.16. (Maximum Risk). Suppose $X_1, \cdots, X_{16}$ are sixteen iid observations from a Bernoulli distribution with parameter $p, 0 < p < 1$. Let $T = \sum_{i=1}^{16} X_i$.

(a) Calculate the risk function of each of the following estimators:

$$\frac{T}{16}; \frac{T+2}{20}.$$

(b) Find the maximum risk of each estimator. Comment.

5.2 References

Berger, J. (2010). Statistical Decision Theory and Bayesian Analysis, Springer, New York.

Bickel, P. and Doksum, K. (2001). Mathematical Statistics, Basic Ideas and Selected Topics, Vol.I, Prentice Hall, NJ

Box, G.E.P. (1953). Nonnormality and tests of variance, Biometrika, 40, 318-335.

Brown, L. (1994). Minimaxity, more or less, In Stat. Dec. Theory and Rel. Topics, S.S. Gupta and J. Berger Eds., Springer, New York.

Brown, L. (2000). An essay on statistical decision theory, JASA, 95, 1277-1281.

DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability, Springer, New York.

Dawid, A.P. (1982). The well calibrated Bayesian, with discussion, JASA, 77, 605-613.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). Robust Statistics: The Approach Based on Influence Functions, Wiley, New York.

Huber, P. (1981). Robust Statistics, Wiley, New York.

Johnstone, I. (2012). Function Estimation, Gaussian Sequence Models, and Wavelet Shrinkage, Cambridge Univ. Press, Cambridge, Forthcomiing.

Kadane, J. and Wolfson, L. (1998). Experiences in elicitation, The Statistician, 47, 1, 3-19.

Lehmann, E. and Casella, G. Theory of Point Estimation, Springer, New York.

Portnoy, S. and He, X. (2000). A robust journey to the new millennium, JASA, 95, 1331-1335.

Seidenfeld, T. (1985). Calibration, coherence, and scoring rules, Phil. Sci., 52, 274-294.

Young, A. and Smith, R. (2010). Essentials of Statistical Inference, Cambridge Univ. Press, Cambridge.

Stigler, S. (2010). The changing history of robustness, Amer. Statist., 64, 277-281.