Chapter 2
Estimating the Mean

This chapter deals with one of the elementary statistical problems, estimating the mean of a random sample from a normal distribution. We assume that the variance of this distribution is known. More general versions of this problem are addressed in later chapters.

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown expectation μ and variance known to be equal to unity. We write $X_i \sim \mathcal{N}(\mu, 1)$, $i = 1, \ldots, n$, independently. Without a careful description of the task related to μ, we would not contemplate any estimator other than the sample mean $\hat{\mu} = (X_1 + \cdots + X_n)/n$. It is unbiased and efficient for μ; its sampling variance is equal to $1/n$. However, if we are averse to positive estimation errors, $\hat{\mu} > \mu$, then an estimator $\hat{\mu} - c$, where c is a positive constant, may be more suitable.

2.1 Estimation with an Asymmetric Loss

Suppose we associate the estimation error $\tilde{\mu} - \mu$ of an estimator $\tilde{\mu}$ of the target μ with loss $(\tilde{\mu} - \mu)^2$ when $\tilde{\mu} < \mu$, but for positive estimation error, when $\tilde{\mu} > \mu$, the loss is $R(\tilde{\mu} - \mu)^2$, where R is a constant greater than unity. This loss, as a function of $\tilde{\mu}$ and μ, is a piecewise quadratic loss function. In fact, the function depends on $\tilde{\mu}$ and μ only through the estimation error $\tilde{\mu} - \mu$. Figure 2.1 displays examples of piecewise quadratic loss functions for two values of the penalty ratio R, each with three values of the target μ. In the left-hand panel, six functions $L(\tilde{\mu}, \mu)$ are drawn, but they correspond to only two distinct functions of the error $\tilde{\mu} - \mu$ in the right-hand panel.

We explore estimators $\tilde{\mu} = \hat{\mu} - c$, where $\hat{\mu}$ is the sample mean and c is a constant that we would set. So, $\tilde{\mu} \sim \mathcal{N}(\mu - c, 1/n)$. The expected loss of $\tilde{\mu}$ is

N. T. Longford, *Statistical Decision Theory*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-40433-7_2,

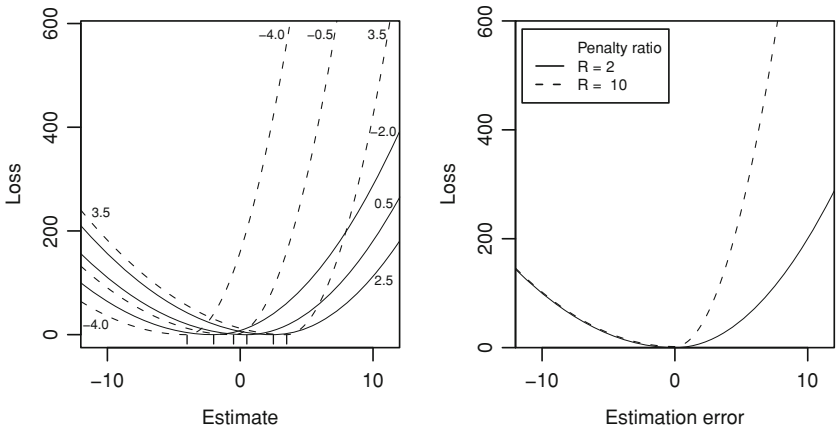

Fig. 2.1 Piecewise quadratic loss functions, as functions of estimate and target *(left-hand panel)* and of the estimation error *(right-hand panel)*. The values of the target μ are indicated in the *left-hand panel*

$$Q = R\sqrt{n}\int_{\mu}^{+\infty}(y-\mu)^2\,\phi\big\{\sqrt{n}(y-\mu+c)\big\}\,\mathrm{d}y + \sqrt{n}\int_{-\infty}^{\mu}(y-\mu)^2\,\phi\big\{\sqrt{n}(y-\mu+c)\big\}\,\mathrm{d}y, \tag{2.1}$$

where ϕ is the density of the standard normal distribution, $\mathcal{N}(0,1)$,

$$\phi(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}y^2\right).$$

Denote by Φ the distribution function of $\mathcal{N}(0,1)$. The transformation $z = \sqrt{n}(y-\mu+c)$ yields the equivalent expression

$$Q = R\int_{c\sqrt{n}}^{+\infty}\left(\frac{z}{\sqrt{n}}-c\right)^2\phi(z)\,\mathrm{d}z + \int_{-\infty}^{c\sqrt{n}}\left(\frac{z}{\sqrt{n}}-c\right)^2\phi(z)\,\mathrm{d}z.$$

Denote the two terms by Q_+ and Q_-. For $R = 1$, we would obtain

$$Q_+ + Q_- = \mathrm{MSE}(\tilde{\mu};\mu) = c^2 + \frac{1}{n},$$

and $c = 0$ would be the optimal choice. For $R \neq 1$, a similar reduction does not take place. We work out the details for Q_-; Q_+ is dealt with similarly. By expanding the square in the integrand, we obtain

$$Q_- = \frac{1}{n}\int_{-\infty}^{c\sqrt{n}} z^2\phi(z)\,\mathrm{d}z - \frac{2c}{\sqrt{n}}\int_{-\infty}^{c\sqrt{n}} z\phi(z)\,\mathrm{d}z + c^2\int_{-\infty}^{c\sqrt{n}} \phi(z)\,\mathrm{d}z\,.$$

It is easy to check that $\phi'(z) = -z\phi(z)$, so $-\phi(z)$ is a primitive function for $z\phi(z)$. The first integral is evaluated by parts, differentiating z and integrating $z\phi(z)$:

$$\begin{aligned} Q_- &= \frac{1}{n}\left\{\left[-z\phi(z)\right]_{-\infty}^{c\sqrt{n}} + \int_{-\infty}^{c\sqrt{n}} \phi(z)\mathrm{d}z\right\} - \frac{2c}{\sqrt{n}}\left[-\phi(z)\right]_{-\infty}^{c\sqrt{n}} + c^2\Phi\left(c\sqrt{n}\right) \\ &= -\frac{c}{\sqrt{n}}\,\phi\left(c\sqrt{n}\right) + \frac{1}{n}\,\Phi\left(c\sqrt{n}\right) + \frac{2c}{\sqrt{n}}\,\phi\left(c\sqrt{n}\right) + c^2\Phi\left(c\sqrt{n}\right) \\ &= \left(c^2 + \frac{1}{n}\right)\Phi\left(c\sqrt{n}\right) + \frac{c}{\sqrt{n}}\,\phi\left(c\sqrt{n}\right). \end{aligned}$$

By similar steps we obtain the identity

$$Q_+ = R\left(c^2 + \frac{1}{n}\right)\left\{1 - \Phi\left(c\sqrt{n}\right)\right\} - \frac{cR}{\sqrt{n}}\,\phi\left(c\sqrt{n}\right). \tag{2.2}$$

Hence the expected loss $Q = Q_+ + Q_-$ is

$$Q = \left(c^2 + \frac{1}{n}\right)\left\{R - (R-1)\,\Phi\left(c\sqrt{n}\right)\right\} - \frac{c(R-1)}{\sqrt{n}}\,\phi\left(c\sqrt{n}\right). \tag{2.3}$$

2.2 Numerical Optimisation

It remains to find the constant c for which the expected loss Q in (2.3) is minimised. This cannot be done by a closed-form expression. We apply the Newton-Raphson algorithm. It is an iterative procedure which generates a provisional (approximate) solution $c^{(i+1)}$ in iteration $i+1$ by adjusting the previous solution $c^{(i)}$. The adjustment depends on the first- and second-order derivatives of Q with respect to c:

$$c^{(i+1)} = c^{(i)} - \frac{s\left(c^{(i)}\right)}{H\left(c^{(i)}\right)}, \tag{2.4}$$

where $s = \partial Q/\partial c$ and $H = \partial^2 Q/\partial c^2$ are treated as functions of c. They are called the score and the Hessian (functions), respectively. The iterations are stopped when the absolute value of the adjustment $c^{(i+1)} - c^{(i)} = -s/H$, or of the score s, becomes smaller than a prescribed small quantity, such as 10^{-8}.

We derive the adjustment (2.4) to gain an understanding of the properties of this algorithm and to formulate its assumptions. Obviously, the first- and second-order derivatives of Q have to exist in the range of plausible values of c. This is satisfied for

Q in (2.3). The Taylor expansion for the first-order derivative at the exact solution c^*, centred around the current (provisional) solution $c^{(i)}$, is

$$s\left(c^*\right) \doteq s\left(c^{(i)}\right) + \left(c^* - c^{(i)}\right) H\left(c^{(i)}\right). \tag{2.5}$$

At the minimum of Q, $s(c^*) = 0$. Regarding (2.5) as an identity, setting aside the fact that it is merely an approximation, and solving it for c^*, we obtain the updating formula in (2.4). Hopefully this gets us closer to c^*. From the derivation, it is clear that this algorithm converges fast when the approximation in (2.5) is precise, that is, when the function s is close to linearity—when Q is close to a quadratic function, or when a solution $c^{(i)}$ is already close to c^*. Problems arise when H is not a smooth function and the values of $1/H$ are not changing at a sedate pace, or indeed when $H = 0$. If iterations reach a region where $H(c) \doteq 0$, the consecutive values $c^{(i)}$ may become unstable. The Newton-Raphson iterations require an initial solution $c^{(0)}$. It can be set by trial and error if we have only one problem to solve. For finding the minimum of Q, $c^{(0)} = 0$ is a suitable initial solution.

When it converges, the Newton-Raphson algorithm finds a root of the score function. The function s may have several roots and the one we find may not be a (global) minimum of Q. However, if s is an increasing function, then the root is unique and it is the only minimum of Q. Often a simple way of proving that s is increasing is by checking that H is positive at the root (or throughout).

For the function in (2.3), we have

$$\begin{aligned} s = \frac{\partial Q}{\partial c} &= 2c\left\{R - (R-1)\,\Phi\left(c\sqrt{n}\right)\right\} - \left(c^2 + \frac{1}{n}\right)\sqrt{n}(R-1)\,\phi\left(c\sqrt{n}\right) \\ &\quad - \frac{R-1}{\sqrt{n}}\,\phi\left(c\sqrt{n}\right) + c^2\sqrt{n}(R-1)\,\phi\left(c\sqrt{n}\right) \\ &= 2cR - 2(R-1)\left\{c\,\Phi\left(c\sqrt{n}\right) + \frac{\phi\left(c\sqrt{n}\right)}{\sqrt{n}}\right\} \\ H = \frac{\partial^2 Q}{\partial c^2} &= 2R - 2(R-1)\left\{\Phi\left(c\sqrt{n}\right) - c\sqrt{n}\,\phi\left(c\sqrt{n}\right) + c\sqrt{n}\,\phi\left(c\sqrt{n}\right)\right\} \\ &= 2\left\{R - (R-1)\,\Phi\left(c\sqrt{n}\right)\right\}. \end{aligned} \tag{2.6}$$

From this we conclude that $2 < H(c) < 2R$ for all $R > 0$ (not only for $R > 1$), so Q has a unique minimum, and it is at the root of s. Since $H(c) > 2$, there are no convergence problems.

By way of an example, suppose $n = 50$ and $R = 20$. We set the initial solution to $c^{(0)} = 0$; the corresponding value of Q is $(R+1)/(2n) = 0.21$. The progression of the provisional solutions is displayed in Table 2.1. The right-most column (Precision) is defined as

$$-\frac{1}{2}\log_{10}\left[\left(c^{(i)} - c^{(i-1)}\right)^2 + \left\{s\left(c^{(i)}\right)\right\}^2\right] \tag{2.7}$$

Table 2.1 Iterations of the Newton-Raphson algorithm to minimise the expected loss Q with penalty ratio $R = 20$ and size $n = 50$ of a random sample from $\mathcal{N}(\mu, 1)$; piecewise quadratic loss

Iteration (i)	$c^{(i)}$	$Q^{(i)}$	Precision
0	0.0000	0.2100	
1	0.1021	0.2100	−0.33
2	0.1511	0.0820	0.27
3	0.1632	0.0674	1.04
4	0.1639	0.0668	2.36
5	0.1639	0.0668	4.93
6	0.1639	0.0668	10.08

(logarithm with base 10). It can be interpreted as the number of digits of precision. The iterations are stopped when this quantity exceeds 8.0. The table indicates that convergence is achieved after six iterations, although we could have stopped after just four. However, the calculations, done in `R`, are instant, so the additional two iterations represent no waste of our resources.

Thus, the estimator with the minimum expected loss when $n = 50$ and $R = 20$ is $\hat{\mu} - 0.1639$ and the corresponding expected loss is 0.0668. The expected loss with the unbiased estimator is 0.2100, more than three times greater. It is easy to show that when $\sigma^2 \neq 1$, $\hat{\mu} - 0.1639\sigma$ is the estimator with the smallest expected loss.

2.3 Plausible Loss Functions

In practice, it is difficult to set the penalty ratio R to a single value without leaving some doubt that R may be somewhat greater or smaller. We regard this as a source of uncertainty associated with the *elicitation* process, the dialogue between the analyst and the client, in which the background and details of the problem are discussed. We address it by solving the problem for a range of values of R that were agreed to be plausible. A range of penalty ratios $(R_{\mathrm{L}}, R_{\mathrm{U}})$, and any value within it, is said to be plausible if the client would rule out any value of R outside this range. At the same time, the plausible range should be set to as narrow an interval as possible.

Figure 2.2 presents the continuum of solutions c^* and the corresponding expected losses for sample sizes $10 \leq n \leq 200$ and penalty ratios $5 \leq R \leq 100$. Denote these functions (curves) by $c_R(n)$ and $Q_R(n)$, respectively. The panels at the top plot c^* and Q as functions of n on the linear (original) scale, and the panels at the bottom reproduce them on the log scale for n. The log scale is useful because at the planning stage one is more likely to consider increasing or reducing the sample size by a certain multiple, such as 1.25 or (25 %), and that corresponds to an increase or reduction by a constant, log (1.25), on the log scale.

The diagram shows that the optimal shift c^* is positive throughout, it increases with R and decreases with n, steeply for small n. For small sample sizes, the functions $c_R(n)$ and $Q_R(n)$ have steep gradients on R, so a lot is at stake. For large sample

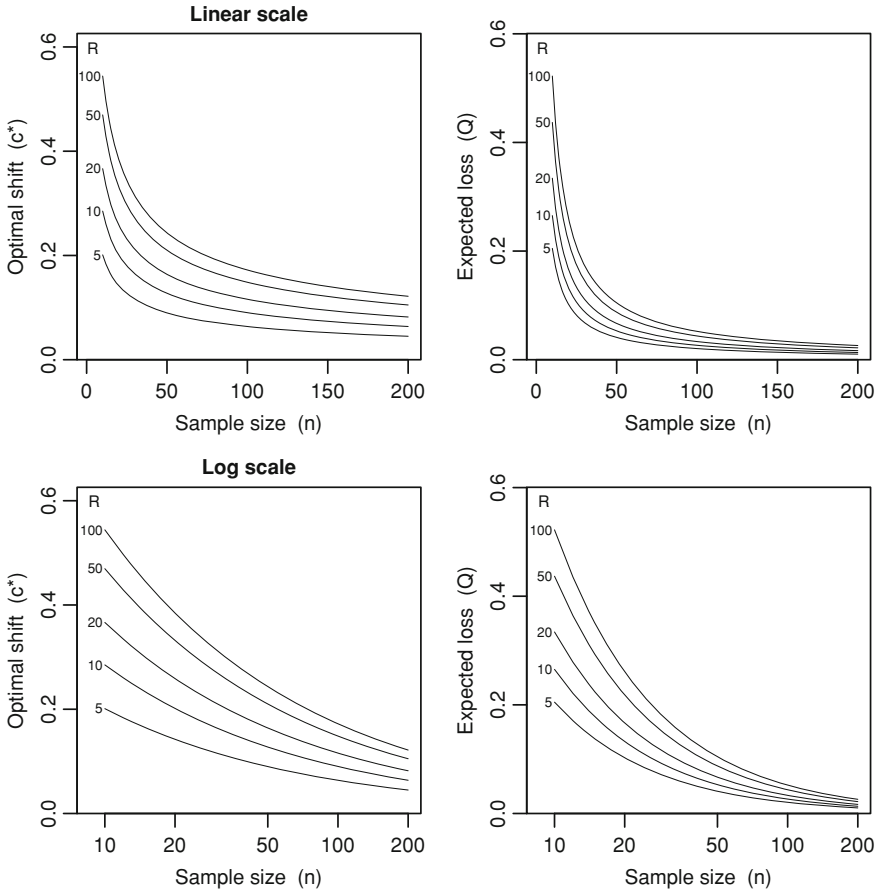

Fig. 2.2 The offset c^* for which the estimator $\hat{\mu} - c^*$ has minimum expected loss, as a function of the sample size n and the penalty ratio R; piecewise quadratic loss. The corresponding expected loss is plotted in the *right-hand panels*

sizes, the differences diminish. In fact, for any fixed R, $c_R(n)$ converges to zero as $n \to +\infty$, but the convergence is rather slow.

Some of these conclusions can be confirmed directly from (2.6). Since H is positive, s is an increasing function of c. But $s(0) = -2(R-1)\phi(0)/\sqrt{n} < 0$, so c^*, the root of s, has to be positive. Further, $s\sqrt{n}$ depends on c and $\sqrt{n}$ only through $c\sqrt{n}$. If c_1 is the root of s for n_1, then $c_2 = c_1\sqrt{n_1/n_2}$ is the root for n_2. Therefore the root of s, which coincides with the root of $s\sqrt{n}$, is a decreasing function of n and $c_R(n) \to 0$ as $n \to +\infty$. Similarly $s\sqrt{n}/(R-1)$ is a decreasing function of R, so long as $c > 0$ and $R > 1$:

$$\frac{s(c)\sqrt{n}}{R-1} = \frac{2c\sqrt{n}R}{R-1} - 2\left\{c\sqrt{n}\,\Phi\left(c\sqrt{n}\right) + \phi\left(c\sqrt{n}\right)\right\}.$$

By definition, $s\{c_R(n)\} = 0$. By substituting $R' > R$ for R and $c_R(n)$ for c on the right-hand side, we obtain a negative quantity. Since s is increasing, $c_{R'}(n)$ has to be greater than $c_R(n)$.

Each curve in Fig. 2.2 is drawn by connecting the values of $c_R(n)$ and $Q_R(n)$ for a fine grid of values n. We set $n = 10, 12, \ldots, 200$. A coarser grid may suffice, but the saving in the computing involved is insubstantial. In fact, all the evaluations for the diagram took only 0.19 sec. of CPU time on a Mac laptop. In `R`, a function is declared, with arguments n, R, and some others that specify the convergence criterion and control the output. One output has the format of Table 2.1, with the details of the iterations, and the other gives only the 'bottom line': c^*, $Q(c^*)$ and the number of iterations. For the evaluations presented in Fig. 2.2, between five and eight iterations were required. The function returns the results for one setting of n and R, but its repeated application for a range of values of n and R requires minimum programming effort, using the system-defined function `apply`.

When the second-order derivative is not available, or we want to avoid its evaluation because it is too complex, the Newton (linearisation) method can be applied. In this method, a pair of provisional solutions, (c_A, c_B), defines the following approximation to the root of s:

$$c_D = c_A - \frac{c_B - c_A}{s(c_B) - s(c_A)}\, s(c_A).$$

This rule is applied iteratively. In the next iteration, the pair (c_B, c_D) is used in place of (c_A, c_B). The iterations are stopped when the two provisional solutions are very close to one another and the value of s for both of them is sufficiently close to zero. A criterion similar to (2.7) can be formulated.

2.4 Other Classes of Loss Functions

In this section, we extend the repertoire of loss functions for which estimation with minimum expected loss is tractable.

The piecewise linear loss for estimator $\hat{\theta}$ of θ is defined as $\theta - \hat{\theta}$ when $\hat{\theta} < \theta$ and as $R(\hat{\theta} - \theta)$ when $\hat{\theta} > \theta$. The penalty ratio $R > 0$ plays a role similar to its namesake for piecewise quadratic loss, to reflect the aversion to positive estimation errors (when $R > 1$). The expected loss of an estimator $\tilde{\mu} = \hat{\mu} - c$ of the mean of the normal distribution with unit variance is

$$R\sqrt{n}\int_{\mu}^{+\infty} (y - \mu)\phi\left\{\sqrt{n}(y - \mu + c)\right\} \mathrm{d}y$$
$$+ \sqrt{n}\int_{-\infty}^{\mu} (\mu - y)\phi\left\{\sqrt{n}(y - \mu + c)\right\} \mathrm{d}y;$$

compare with (2.1). By steps similar to those used in deriving (2.3), we obtain

$$Q = \frac{R}{\sqrt{n}}\Big[-\phi(z)\Big]_{c\sqrt{n}}^{+\infty} - cR\left\{1 - \Phi\big(c\sqrt{n}\big)\right\} + c\,\Phi\big(c\sqrt{n}\big) - \frac{1}{\sqrt{n}}\Big[-\phi(z)\Big]_{-\infty}^{c\sqrt{n}}$$
$$= (R+1)\left\{c\,\Phi\big(c\sqrt{n}\big) + \frac{\phi\big(c\sqrt{n}\big)}{\sqrt{n}}\right\} - cR.$$

Its derivatives with respect to c are

$$s = (R+1)\,\Phi\big(c\sqrt{n}\big) - R$$
$$H = (R+1)\sqrt{n}\,\phi\big(c\sqrt{n}\big)\,,$$

simpler than for the piecewise quadratic loss. We have a closed-form solution for minimising Q,

$$c^* = \frac{1}{\sqrt{n}}\,\Phi^{-1}\left(\frac{R}{R+1}\right). \tag{2.8}$$

The discussion of the properties of this solution is left for an exercise.

In principle, any loss function can be declared that is increasing in the absolute estimation error $\Delta = |\,\tilde{\mu} - \mu\,|$ and for which $L(0) = 0$. The latter condition is not important, because we could adjust L as $L - L(0)$; we only need $L(0)$ to be well defined. A reasonable condition is that L be continuous, although it does not have to be differentiable throughout. Apart from additivity (Sect. 1.7), the key criterion of usefulness of a loss function is that it reflects the client's perspective. The following example shows, however, that some loss functions lead to unreasonable answers.

The piecewise constant loss function is defined as the constant unity for negative estimation error and $R > 0$ for positive estimation error. The expected loss for estimating μ by $\tilde{\mu} = \hat{\mu} - c$ is

$$Q = R\sqrt{n}\int_{\mu}^{+\infty}\phi\big\{\sqrt{n}(y - \mu + c)\big\}\,\mathrm{d}y + \sqrt{n}\int_{-\infty}^{\mu}\phi\big\{\sqrt{n}(y - \mu + c)\big\}\,\mathrm{d}y$$
$$= R - (R-1)\,\Phi\big(c\sqrt{n}\big)\,,$$

which has no minimum, but suggests the solution $c^* = +\infty$, that is, the 'estimator' $\tilde{\mu} = -\infty$. Since $\mathrm{P}(\tilde{\mu} \neq \mu) = 1$, we are certain to pay a penalty. We should therefore avoid positive estimation error (with penalty $R > 1$), and that is achieved with a sufficiently small estimate (large c). Thus, reducing our attention to the sign of the estimation error is a bad strategy; its size also matters.

2.4.1 LINEX Loss

The LINEX loss for estimation error $\Delta = \hat{\theta} - \theta$ is defined as

$$L_a(\Delta) = \exp(a\Delta) - a\Delta - 1;$$

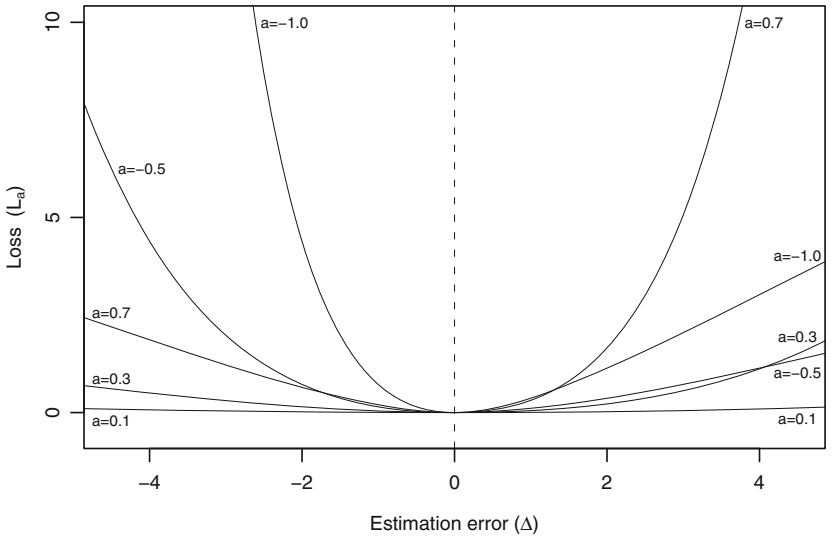

Fig. 2.3 Examples of LINEX loss functions

$a \neq 0$ is a constant to be set. It is easy to check that L_a has all the attributes of a loss function: $L_a(0) = 0$ and $L_a(\Delta)$ decreases for negative Δ and increases for positive Δ. The function is drawn in Fig. 2.3 for a few coefficients a.

For $x \gg 0$ (x positive and large), $\exp(x) \gg x+1$, so when a and Δ have the same sign and $a\Delta$ is large, $(a\Delta+1)/L_a(\Delta) \doteq 0$ and $L_a(\Delta)$ behaves similarly to $\exp(a\Delta)$. In contrast, when a and Δ have opposite signs and $-a\Delta$ is large, $\exp(a\Delta)/L_a(\Delta) \doteq 0$, and $L_a(\Delta) \ll 1 - a\Delta \ll L_a(-\Delta)$. So, L_a is distinctly asymmetric, with greater values for large positive errors when $a > 0$, and greater values for large negative errors when $a < 0$.

The expected LINEX loss of $\hat{\mu} - c$ is

$$\begin{aligned} Q_a &= \sqrt{n} \int_{-\infty}^{+\infty} L_a(y-\mu)\, \phi\big\{\sqrt{n}\,(y-\mu+c)\big\}\, \mathrm{d}y \\ &= \int_{-\infty}^{+\infty} \exp\left(\frac{az}{\sqrt{n}} - ac\right) \phi(z)\, \mathrm{d}z \;-\; \int_{-\infty}^{+\infty} \left(\frac{az}{\sqrt{n}} - ac\right) \phi(z)\, \mathrm{d}z - 1. \end{aligned}$$

The latter integral is equal to $-ac$ because ϕ is symmetric and it integrates to unity. The former integral can be related to the expectation of a lognormal distribution. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathrm{E}\{\exp(X)\} = \exp(\mu + \frac{1}{2}\sigma^2)$. To make the text self-contained, we derive it from basic principles.

By consolidating the arguments of the exponentials and matching the result with a normal density, $\mathcal{N}(a/\sqrt{n}, 1)$, we obtain

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} \exp\left(\frac{az}{\sqrt{n}} - ac\right) \phi(z)\,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \exp(-ac) \int_{-\infty}^{+\infty} \exp\left(-\frac{z^2}{2} + \frac{az}{\sqrt{n}}\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{a^2}{2n} - ac\right) \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\left(z - \frac{a}{\sqrt{n}}\right)^2\right\} \mathrm{d}z \\
&= \exp\left(\frac{a^2}{2n} - ac\right).
\end{aligned}
$$

Hence

$$
Q_a \;=\; \exp\left(\frac{a^2}{2n} - ac\right) + ac - 1.
$$

The minimum of this function of c is found by exploring its derivative:

$$
s_a \;=\; a\left\{1 - \exp\left(\frac{a^2}{2n} - ac\right)\right\}.
$$

Further differentiation yields the Hessian

$$
H_a \;=\; a^2 \exp\left(\frac{a^2}{2n} - ac\right).
$$

Since $H_a > 0$, Q_a has a unique minimum, and it is at the root of s_a. The root is $c^* = a/(2n)$ and the minimum attained is $Q_a(c^*) = ac^* = a^2/(2n)$. The expected loss with $\hat{\mu}$, which corresponds to $c = 0$, is $\exp\{a^2/(2n)\} - 1$. The difference of the losses, $\exp\{a^2/(2n)\} - a^2/(2n) - 1$, is equal to the loss $L_1\{a^2/(2n)\} = L_a\{a/(2n)\}$. The expected loss decreases with n to zero, but for small to moderate n it is substantial, especially when $|a|$ is large.

2.5 Comparing Two Means

In this section we address the problem of deciding which of two random samples from normal distributions with the identical variances is greater. We assume that the common variance, σ^2, is known. No generality is lost by assuming that $\sigma^2 = 1$, because we can reformulate the problem for samples $\mathbf{x}_1$ and $\mathbf{x}_2$ as a problem for $\sigma^{-1}\mathbf{x}_1$ and $\sigma^{-1}\mathbf{x}_2$. Denote the expectations of the two samples by μ_1 and μ_2 and set $\Delta = \mu_2 - \mu_1$. Let n_1 and n_2 be the sizes of the two samples and $\widehat{\Delta} = \hat{\mu}_2 - \hat{\mu}_1$ the difference of the sample means. Its distribution is $\mathcal{N}(\Delta, m\sigma^2)$, where $m = 1/n_1 + 1/n_2$; $1/m$ can be interpreted as the effective sample size of the pair of the samples.

With hypothesis testing, we set the size of the test, α, by convention to 0.05, although other choices (probabilities) are permitted, and choose the critical region, denoted by $\mathcal{C}$, such that under the (null) hypothesis that $\Delta = 0$ the probability that a new realisation of $\widehat{\Delta}$ falls within $\mathcal{C}$ is equal to α. Common choices for $\mathcal{C}$ are the complement of a symmetric interval, $\{-\infty, \sigma\sqrt{m}\,\Phi^{-1}(\frac{1}{2}\alpha)\} \cup \{\sigma\sqrt{m}\,\Phi^{-1}(1-\frac{1}{2}\alpha), +\infty\}$, and the one-sided intervals $\{\sigma\sqrt{m}\,\Phi^{-1}(1-\alpha), +\infty\}$ and $\{-\infty, \sigma\sqrt{m}\,\Phi^{-1}(\alpha)\}$. If $\widehat{\Delta} \in \mathcal{C}$, we reject the null hypothesis. Otherwise, we have no evidence against the null hypothesis. Interpreting the latter outcome as a confirmation that $\Delta = 0$, or that $|\Delta|$ is small, is not appropriate. Following it up by action that would be apropriate if $\Delta = 0$ but not otherwise, has no logical basis.

Suppose we have a research or business agenda the details of which depend on Δ. If we knew that $\Delta < 0$, action A would be appropriate. Otherwise we would pursue action B. If we elect action A but $\Delta > 0$, we incur loss μ^2; if we elect action B but $\Delta < 0$, we lose $R\mu^2$. Note that this loss function differs from the function of the same name defined in Sect. 2.1, because no loss is incurred when the correct sign is chosen, even if $\widehat{\Delta}$, or another estimate, differs substantially from Δ. Because of the symmetry of the problem, we lose no generality by assuming that $R > 1$, so that its label, *penalty* ratio, is well motivated.

We intend to base the choice between A and B on $\widehat{\Delta} - c$, where c is a constant to be set by the criterion of minimum expected loss. Since $(\widehat{\Delta} - \Delta)/\sqrt{m}$ has the standard normal distribution, $\mathcal{N}(0, 1)$, we can represent Δ by a random variable $\widehat{\Delta}+\delta$, where $\delta \sim \mathcal{N}(0, m)$. Note that we rely in this on the symmetry of $\mathcal{N}(0, m)$. Thus, we have converted an unknown constant, Δ, into a random variable, to represent our uncertainty about its value after its estimate $\widehat{\Delta}$ has been realised; that is, we converted it from a random variable to a constant. We will make these changes of status more formal in Chap. 4 within a Bayesian perspective.

When $\widehat{\Delta} - c < 0$, and so we choose action A, the expected loss is

$$\begin{aligned} Q_- &= \frac{1}{\sqrt{m}} \int_0^{+\infty} x^2 \phi\left(\frac{x - \widehat{\Delta}}{\sqrt{m}}\right) \mathrm{d}x \\ &= \int_{-a}^{+\infty} \left(\widehat{\Delta} + z\sqrt{m}\right)^2 \phi(z)\, \mathrm{d}z \\ &= \widehat{\Delta}^2 \{1 - \Phi(-a)\} - 2\widehat{\Delta}\sqrt{m}\left[\phi(z)\right]_{-a}^{+\infty} + m \int_{-a}^{+\infty} z^2 \phi(z)\, \mathrm{d}z\,, \end{aligned}$$

where $a = \widehat{\Delta}/\sqrt{m}$. The latter integral is evaluated by parts,

$$\int_{-a}^{+\infty} z^2 \phi(z)\, \mathrm{d}z \;=\; \Phi(a) - a\phi(a),$$

exploiting the symmetry of the standard normal distribution, that is, $\phi(-a) = \phi(a)$ and $1 - \Phi(-a) = \Phi(a)$. Therefore,

$$Q_- = m\left\{\left(1+a^2\right)\Phi(a) + a\phi(a)\right\}.$$

By similar steps we obtain the expected loss when choosing action B:

$$Q_+ = mR\left[\left(1+a^2\right)\{1-\Phi(a)\} - a\phi(a)\right].$$

For $\widehat{\Delta}$ given, we select the action with the smaller expected loss. For small values of $\widehat{\Delta}$ action A and for large values action B is preferred. There is a critical value of $\widehat{\Delta}$ where we switch from the preference for one action to the other. This occurs at the *equilibrium*, where $Q_- = Q_+$. To prove that there is a unique equilibrium, we show that Q_- is increasing and Q_+ is decreasing. The derivatives of these functions of a are

$$\begin{aligned}\frac{\partial Q_-}{\partial a} &= 2m\{a\Phi(a) + \phi(a)\}\\ \frac{\partial Q_+}{\partial a} &= 2mR\left[a\{1-\Phi(a)\} - \phi(a)\right].\end{aligned}$$

Both derivatives, as functions of a, are increasing because their respective derivatives, $2m\Phi(a)$ and $2mR\{1-\Phi(a)\}$, are positive. Whereas $\partial Q_-/\partial a$ is positive, because its limits at $\pm\infty$ are zero and $+\infty$, $\partial Q_+/\partial a < 0$, because its limits are $-\infty$ and zero. Therefore Q_- is increasing and Q_+ is decreasing throughout $(-\infty, +\infty)$. So, our best bet is to set the constant c at the equilibrium, where $Q_- = Q_+$. This condition is

$$\Delta Q = (R+1)\left\{\left(1+a^2\right)\Phi(a) + a\,\phi(a)\right\} - R(1+a^2) = 0, \tag{2.9}$$

with the factor m dropped. It is solved by the Newton-Raphson algorithm, using the expression

$$\frac{\partial \Delta Q}{\partial a} = 2(R+1)\{a\,\Phi(a) + \phi(a)\} - 2aR.$$

For the solution a^*, the optimal constant c is $c^* = a^*\sqrt{m}$. The decision about the sign of Δ is based on the sign of $\widehat{\Delta} - a^*\sqrt{m}$. It is rather fortuitous that (2.9) involves the sample sizes n_1 and n_2 only through m, and m only through a. Therefore, it is practical to solve (2.9) for a range of values of R, and then convert the solution a_R^* to $c_R = a_R^*\sqrt{m}$.

For the piecewise linear loss, we find the equilibrium by evaluating the two parts of the expected loss:

$$\begin{aligned}Q_- &= \frac{1}{\sqrt{m}}\int_0^{+\infty} x\,\phi\left(\frac{x-\widehat{\Delta}}{\sqrt{m}}\right)\mathrm{d}x\\ &= \widehat{\Delta}\{1-\Phi(-a)\} - \Big[\phi(z)\Big]_{-a}^{+\infty}\\ &= \sqrt{m}\{a\Phi(a) + \phi(a)\}\end{aligned}$$

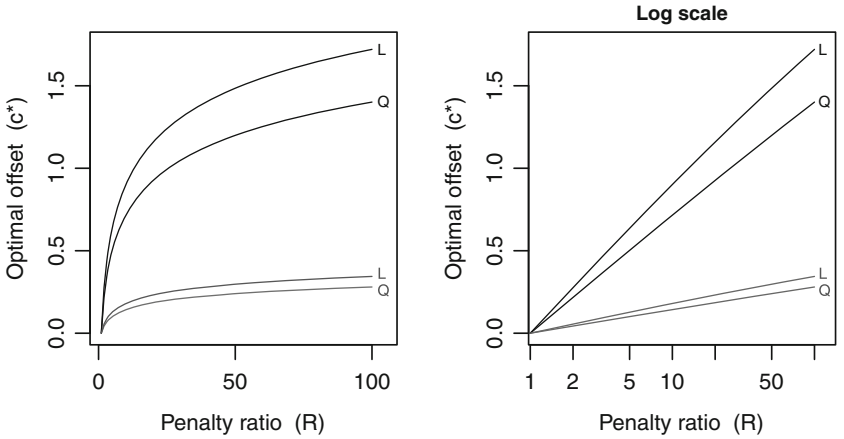

Fig. 2.4 The optimal offsets c_R with the quadratic (Q) and linear loss (L) for $m = 1$ *(black)* and $m = 1/5$ *(gray)*, as functions of the penalty ratio R, on the linear and log scales

and

$$Q_+ = R\sqrt{m}\,\{a\Phi(a) + \phi(a) - a\}.$$

Hence the balance equation

$$\Delta Q = (R-1)\,\{a\Phi(a) + \phi(a)\} - Ra = 0,$$

which is solved by the Newton-Raphson algorithm, in which we use the identity $\partial \Delta Q/\partial a = (R-1)\Phi(a) - R$.

Figure 2.4 displays the solutions c_R for the linear and quadratic loss functions with penalty ratios $R \in (1, 100)$ for $m = 1$ (e.g., $n_1 = n_2 = 2$, drawn in black) and $m = 0.2$ (e.g., $n_1 = n_2 = 10$, gray). The function c_R increases with R, approximately linearly on the log scale (see the right-hand panel). Of course, c^* is smaller for larger samples, in proportion of $\sqrt{m}$. Piecewise linear and quadratic loss functions are, strictly speaking, not comparable even when defined with the same penalty ratio R. However, a client may not be certain as to which of these loss functions is appropriate, so contemplating both of them is within the spirit of a wide range of plausible loss functions.

2.6 Problems, Exercises and Suggested Reading

1. Compare by simulations the sampling variances of the mean, median, the average of the two quartiles and the average of the minimum and maximum of a simple random sample from a normal distribution. Repeat this exercise with the uniform distribution on $(0, 2\theta)$ to estimate θ.

2. Derive in detail the identity in (2.2).
3. Discuss methods for finding the root of s without using its derivative H. Compare the programming effort, the results and the speed of convergence with the Newton-Raphson algorithm on examples of your choice.
4. Discuss the properties of c^* in (2.8). How is c^* adjusted when the variance σ^2 is different from unity? Compare the minimum value of Q with its value for $c = 0$. Study their difference and ratio as $n \to +\infty$.
5. Plot the values of the optimal shift $c_R(n)$ as functions of R for a selection of sample sizes n. Explore the function `contour` in `R` and apply it to the values of $c_R(n)$.
6. Discuss the advantages of working with n and R on the multiplicative scale.
7. Construct a loss function of your own choice based on the properties you would find desirable for a specific example or application. Search the literature for examples of loss functions and discuss their properties. Plot these loss functions, e.g., using the layout of Fig. 2.1.
8. Discuss how the loss function should be adapted for estimating a transformed parameter. For example, we may have a particular loss function for estimating the mean μ of a normal random sample (with known variance), but we wish to know the value of $\exp(\mu)$. Suggested reading about the lognormal distribution: Aitchison and Brown (1957) and Crow and Shimizu (1988). See Longford (2009) for estimating the mean and median of the lognormal distribution in small samples.
9. Show that loss functions form classes of equivalence. Two loss functions, L_1 and L_2, fall into the same class if $L_1 = bL_2$ for some scalar $b > 0$. When is a linear combination of two loss functions, $aL_1 + bL_2$, also a loss function? Construct such a loss function.
10. Discuss how the results of Sect. 2.5 can be applied to deciding whether the expectation of a normally distributed sample (with a known variance) is positive or negative.
    Hint: Suppose in the comparison of two samples, one is so large that its expectation is, in effect, known.
11. Explore estimation of μ in the context of Sect. 2.5 with piecewise constant loss.
12. The switch between the statuses of fixed and random for the parameter of interest in Sect. 2.5 is associated with the fiducial argument. See Seidenfeld (1992) for background.
13. Suggested reading about methods for optimisation: Lange (1999), Chaps. 5, 11, 13, and Press et al. (2007), Chap. 9.
14. Suggested reading about LINEX loss: Zellner (1986).
15. Suggested reading of historical interest: Friedman and Savage (1948); Wald (1950); Le Cam (1955); Pratt et al. (1964) . Also, Pratt et al. (1995).
16. An unsolved problem. Why is c^* approximately linearly related to $\log(R)$?
17. Derive the integral of Φ and the integral of the result, and explain the appearance of these functions in Sect. 2.4.

References

Aitchison, J., & Brown, J. A. C. (1957). *The lognormal distribution*. Cambridge: Cambridge University Press.

Crow, E. L., & Shimizu, K. (Eds.). (1988). *Lognormal distributions*. New York: Theory and Applications. M. Dekker.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*, 279–304.

Lange, K. (1999). *Numerical analysis for statisticians*. New York: Springer-Verlag.

Le Cam, L. (1955). An extension of Wald's theory of statistical decision functions. *Annals of Mathematical Statistics*, *26*, 69–81.

Longford, N. T. (2009). Inference with the lognormal distribution. *Journal of Statistical Planning and Inference*, *139*, 2329–2340.

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1964). The foundations of decision under uncertainty: An elementary exposition. *Journal of the American Statistical Association*, *59*, 353–375.

Pratt, J. W., Raiffa, H., & Schlaifer, R. (1995). *Introduction to statistical decision theory*. Cambridge: MIT Press.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (2007). *Numerical recipes: The art of scientific computing (3rd ed.)*. Cambridge University Press, New York.

Seidenfeld, T. (1992). R. A. Fisher's fiducial argument and Bayes' theorem. *Statistical Science*, *7*, 358–368.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, *81*, 446–451.

http://www.springer.com/978-3-642-40432-0

Statistical Decision Theory
Longford, N.T.
2013, X, 124 p. 23 illus., Softcover
ISBN: 978-3-642-40432-0