Some decision-theoretic aspects of sampling

Yosef Rinott [a][*]

[a]Department of Statistics
and Center for the Study of Rationality
The Hebrew University, Jerusalem 91905, Israel

This chapter offers a glimpse into the literature on finite population sampling from a decision-theoretic point of view. Starting from a formal description of the problem of estimation in finite populations, we present and discuss a selection of result and proofs on optimality of sampling and estimation strategies.

1. Introduction

Decision theory provides tools and insights for understanding, comparing, and selecting sampling and estimation procedures. In this chapter we present a small sample of the extensive literature on decision-theoretic aspects of sampling from finite populations, without attempting to give a comprehensive survey of the best possible results and references.[1] Technical details are sometimes omitted for the sake of simplicity.

The chapter is quite theoretical, dealing with the foundations of *finite population sampling* and inference through simple designs and models rather than the complex ones in modern use. It is hoped that a practitioner may find these basic ideas of interest, albeit theoretical. On the other hand, it seems that a student or teacher of statistical decision theory can definitely benefit from the wealth of ideas that exist in the area of finite population sampling. It provides setups and examples that add an interesting perspective to the standard illustrations given in most statistical decision theory courses, where a *sample* is often restricted to mean i.i.d. observations.

The task of estimating the mean, say, of a given finite population of size N by measuring $n < N$ units does not seem to involve any probability structure, unlike other statistical setups where it is assumed at the outset that the data consist of random or noisy observations. By random sampling, statisticians introduce noise or randomness that did not

---

[1]For a scholarly survey of results until 1987 and numerous references, see Chaudhuri and Vos (1988).

exist in the original problem. It is well known that the introduction of random sampling can avoid biases and allow important notions such as *unbiased estimation* and *confidence intervals*. While many statisticians (and most standard books on sampling) take random sampling as so self-evident that questions like "why do statisticians use dice or other random devices and add randomness or noise to the task" seem unwarranted[2], it is, in fact, an intriguing question that merits more than intuitive answers. Indeed, there is a large body of literature showing formally and precisely that certain relevant optimality criteria can only be achieved by random sampling designs.

Emphasis in this chapter is placed on optimal inference. In the context of finite populations, optimality is most often expressed in terms of minimax results, which in general require random strategies. Other decision-theoretic notions such as loss and risk, admissibility, sufficiency, completeness, unbiasedness, uniformly minimum variance (UMV), Bayes procedures, and more, will also be discussed in connection with finite population sampling.

2. Notations and definitions

The following notation will be used throughout the chapter. A list of main notations appears in Section 8.

1. The ***population*** $\mathcal{Y} = (y_1, \ldots, y_N)$ is a vector of values of some measurements with index set $\mathcal{N} = \{1, \ldots, N\}$, where the population size N is assumed to be known whenever it is needed. Here $i \in \mathcal{N}$ denotes the ***label*** of the i-th population ***unit*** whose value is y_i. In this chapter we assume that $\mathcal{Y} \in \mathbb{R}^N$, so that each y_i is a univariate measurement (although in many applications more than one variable is measured for each unit). Some of the ideas could be extended to more general measurements, but this will not be done here. $\mathcal{Y}$ is an unknown ***parameter***, and so is any function $\theta(\mathcal{Y})$ such as $\bar{\mathcal{Y}} = \frac{1}{N} \sum_{i=1}^{N} y_i$, $V(\mathcal{Y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{\mathcal{Y}})^2$, $Max(\mathcal{Y}) = \text{Max}_{1 \leq i \leq N} y_i$, or $Med(\mathcal{Y}) = \text{Median}_{1 \leq i \leq N} y_i$.

   The set of possible $\mathcal{Y}$'s is denoted by Υ, the ***parameter space***; unless otherwise stated (towards the end of Section 6), we shall **always** assume that Υ is a ***symmetric parameter space***, that is, a symmetric subset of $\mathbb{R}^N$ in the sense that if $\mathcal{Y} = (y_1, \ldots, y_N) \in \Upsilon$, then so does every permutation of $\mathcal{Y}$. In particular, any set of the form $\Upsilon = \Lambda \times \cdots \times \Lambda$, a product of some set N times, satisfies this assumption. The set $\Omega(\mathcal{Y})$ of all permutations of a given vector $\mathcal{Y} = (y_1, \ldots, y_N)$ is of course also symmetric. As usual, the parameter space Υ is known to the statistician.

---

[2]but see Valliant, Dorfman, and Royall (2000) for a refreshing change.

If the parameter $\theta(\mathcal{Y})$ remains constant under permutations of $\mathcal{Y}$, we say that it is a ***symmetric parameter***. The above examples are all of this kind.

2. A ***sampling design*** $\mathcal{P}$ is a probability function on the space of all subsets S of $\mathcal{N}$. Unless otherwise stated we assume ***noninformative sampling***, also known as ***ignorable sampling***; that is, the probability $\mathcal{P}(S)$ does not depend on the parameter $\mathcal{Y}$. Formally, $\mathcal{P}(S\,|\,\mathcal{Y}) = \mathcal{P}(S)$. In the Bayesian or superpopulation context of Section 6, $\mathcal{Y}$ is also random, $\mathcal{P}(S\,|\,\mathcal{Y})$ becomes a conditional probability, and ignorability is equivalent to independence of S and $\mathcal{Y}$.

In certain example we allow the design $\mathcal{P}$ to depend on known covariates, or auxiliary variables; see below. The ***inclusion probability*** of a unit is defined by $\alpha_i = \mathcal{P}(\{i \in S\}) = \sum_{S:S \ni i} \mathcal{P}(S)$, the probability that unit i is in the sample S. Here S is the ***set*** of drawn labels (without order and repetitions). By a simple sufficiency argument given in Remark 1 below, we can ignore designs that take an order of the elements in the sampled set into account or allow repetitions.

The set S is called the ***sample***, and its size, $|S|$, is the ***sample size***. If $\mathcal{P}(S) > 0$ implies $|S| = n$, then the design $\mathcal{P}$ is said to have a ***fixed sample size***.

Simple random sampling without replacement of size n, abbreviated **SRS**, is denoted by $\mathcal{P}_s$ and satisfies $\mathcal{P}_s(S) = 1/\binom{N}{n}$ if $|S| = n$, and zero otherwise.

When ***auxiliary*** information is available in the form of positive values $(x_1, \ldots, x_N)$, where x_i is some *known* value of a variable pertaining to unit $i \in \mathcal{N}$, it can be used in the design and in estimation. For example, when $x_i > 0$, the design having a fixed sample size n, defined by $\mathcal{P}_{ppas}(S) = \sum_{i \in S} x_i / \left[N\bar{\mathcal{X}} \binom{N-1}{n-1} \right]$ if $|S| = n$, where $\bar{\mathcal{X}} = \frac{1}{N}\sum_{i=1}^{N} x_i$, is of this kind (Lahiri, 1951). The notation *ppas* stands for ***probability proportional to aggregate size***, in this case, to the aggregate size of the auxiliary variables in S, $\sum_{i \in S} x_i$. It can be implemented by first choosing one unit from the population, say i, with probability $x_i / N\bar{\mathcal{X}}$, and then adding a subset of $n - 1$ additional units chosen from the remaining $N - 1$ units uniformly, that is, with equal probabilities for for all subsets of size $n - 1$. See Rao and Vijayan (1977) and Hedayat and Sinha (1991) for details and references on this design and a discussion of drawing mechanisms for design implementation, and Cassel, Särndal, and Wretman (1977) for further references.

3. The ***data*** consist of the set of pairs $\{(i, y_i) : i \in S\}$, that is, the y-values and their labels for the units in the sample S. we set

$$D = D[S, \mathcal{Y}] = \{(i, y_i) \,:\, i \in S\}. \tag{1}$$

For $S = \{i_1, \ldots, i_n\}$, let $\mathcal{Y}_S$ be the **multiset** $\{y_{i_1}, \ldots, y_{i_n}\}$, with equal y-values listed separately provided that they have different labels. In other words, $\mathcal{Y}_S$ can be viewed as the sequence $(y_{i_1}, \ldots, y_{i_n})$, where the order is ignored. For example, if $S = \{1, 2, 3\}$ and $y_1 = y_2 = 13$ and $y_3 = 7$, then $\mathcal{Y}_S = \{13, 13, 7\}$ in any order.

The alternative notation $D = D[S, \mathcal{Y}_S]$ indicates that the data consist of S and the coordinates of $\mathcal{Y}$ having labels in S.

Remark 1 *By sufficiency arguments (Basu 1958) we shall consider the data D as above, i.e., without taking into account the order (if known) in which the sample was drawn; when the sampling procedure allows repetitions of units, as in sampling with replacement, repetitions will also be ignored and each repeated unit will be counted once. Since the relevant data D consist only of the set of drawn labels S and their y-values, we shall only consider* **designs $\mathcal{P}$ on the space of (unordered) subsets (with no repetitions)** *of $\mathcal{N}$. The sufficiency of D is intuitively obvious: no information is added by measuring a unit more than once, or specifying the order in which the measurements were taken. A formal statement and proof follow. We denoted designs which ignore the order of labels and repetitions by $\mathcal{P}$, and the corresponding data by $\mathcal{D}$. In the proposition below we consider designs that are probability measures on ordered multisets of $\mathcal{N}$, so repetitions are allowed, and the data contain information on order and repetitions. In this case the sampling design and data are denoted by bold-face letters $\mathbf{P}$ and $\mathbf{D}$, respectively, and the sample is an ordered multiset (allowing repetitions) denoted by $\mathbf{S}$, distributed according to $\mathbf{P}$.*

Proposition 2 *Let $\mathbf{P}$ be a sampling design on* **ordered multisets** *which we denote by $\mathbf{S}$, and consider the data $\mathbf{D} = \{(i, y_i) \ : \ i \in \mathbf{S}\}$, a multiset that includes information on the order and repetitions in the sample. Let $S = r(\mathbf{S}) = \{i : i \in \mathbf{S}\}$; that is, S is the set formed from $\mathbf{S}$ when repetitions and order are ignored, and let $D = r(\mathbf{D}) = \{(i, y_i) \ : \ i \in S\}$. Then D is a* **sufficient statistic** *for the parameter $\mathcal{Y}$.*

Proof *For a design $\mathbf{P}$ as above, the conditional probability of $\mathbf{D} = \{(i, y_i) \ : \ i \in \mathbf{S}\}$ given D, where $\mathbf{S}$ is an ordered multiset and the parameter is $\mathcal{Y}$, satisfies*

$$P(\mathbf{D}|D) = \begin{cases} \mathbf{P}(\mathbf{S}) / \displaystyle\sum_{\mathbf{S}': r(\mathbf{S}')=S} \mathbf{P}(\mathbf{S}') & if \quad r(\mathbf{D}) = D \\ \\ 0 & otherwise. \end{cases} \tag{2}$$

Since the right-hand side of (2) depends only on D and not on the parameter, it follows that D is sufficient. ∎

4. An **estimator** $t = t(D) = t(\{(i, y_i) : i \in S\})$ is a function of the data. We use various notations for $t(D)$, namely $t(D[(S, \mathcal{Y}_S)])$, or $t(S, \mathcal{Y}_S)$, or $t(S, \mathcal{Y})$; the first two emphasize that t depends on $\mathcal{Y}$ only through values with labels in S, and the third is less cumbersome and sometimes more convenient. It should be emphasized that $t(S, \mathcal{Y})$ depends only on the data $\{(i, y_i) : i \in S\}$, that is, the labels and the labeled y-values in the sample. Note that when the sample size $|S|$ is not fixed, then implicit in the notation is the assumption that t is a function defined on arguments of different dimensions.

   If the estimator $t(S, \mathcal{Y}) = t(D[(S, \mathcal{Y}_S)])$ can be expressed as a function of $\mathcal{Y}_S$ alone, we write $t(S, \mathcal{Y}) = t(\mathcal{Y}_S)$ and say that t is **symmetric** (or invariant). Such an estimator depends on the y-values in the sample, and not on their labels.

   Examples of symmetric statistics are the sample mean $\bar{y}_S = \frac{1}{|S|} \sum_{i \in S} y_i$ and variance $\frac{1}{|S|-1} \sum_{i \in S} (y_i - \bar{y}_S)^2$. On the other hand, the **Horvitz-Thompson estimator** $t_{HT} = \sum_{i \in S} y_i / \alpha_i$ does require knowledge of the labels associated with each y-value and thus it is **not** symmetric.

   When **auxiliary** information $(x_1, \ldots, x_N)$ is available for every unit in the population, it can be used in the sampling design and in estimation. For example, consider the **ratio estimator** of $\bar{\mathcal{Y}}$ defined by $t_R = (\bar{y}_S / \bar{x}_S)\bar{\mathcal{X}}$, where $\bar{x}_S = \frac{1}{|S|} \sum_{i \in S} x_i$; we denote it by t_R since $\bar{y}_S / \bar{x}_S$ is an estimator of the ratio $R = \bar{\mathcal{Y}} / \bar{\mathcal{X}}$. The estimator t_R is **not** symmetric since the computation of $\bar{x}_S$ requires knowledge of the labels in S (but not their pairing with the y-values). Note that if $\bar{\mathcal{X}}$ is known, and the population consists of the pairs, that is, $\mathcal{Z} = \{z_1 = (y_1, x_1), \ldots, z_N = (y_N, x_N)\}$, then t_R is a symmetric estimator for the population $\mathcal{Z}$.

   In this chapter we assume $t \in \mathbb{R}$ and any value in $\mathbb{R}$ is allowed, regardless of the parameter space. For example, a proportion in a population of size N (see Section 3.5.2) is necessarily a rational number of the form k/N, but we allow an estimator t of this proportion to assume any real value; if certain values are undesired, the loss function should reflect it.

5. A pair $(\mathcal{P}, t)$ consisting of a sampling design and an estimator is called a **strategy**. A **class of strategies** consists of all pairs $(\mathcal{P}, t)$ such that $\mathcal{P}$ belongs to some class of sampling designs and t belongs to some class of estimators.

6. A **loss function** $L(\tau, \mathcal{Y})$ represents a penalty paid in an estimation problem when the estimator assumes the value τ, and the value of the parameter is $\mathcal{Y}$. If $\theta = \theta(\mathcal{Y})$ is a parameter and $t = t(S, \mathcal{Y}_S)$ is an estimator of θ, we may use the notation $L(t, \theta)$ for the loss. A common example is $L(t, \theta) = (t - \theta)^2$, the **quadratic** loss function (= squared error loss). A loss function is said to be **symmetric** if $L(\tau, \mathcal{Y})$ remains constant when $\mathcal{Y}$ is replaced by any permutation of its coordinates for any fixed τ. Clearly, if $\theta(\mathcal{Y})$ is a **symmetric parameter**, that is, if it remains constant under permutations of $\mathcal{Y}$, then so does $L(\tau, \theta(\mathcal{Y}))$, and the loss is symmetric.

7. The **risk** of a strategy $(\mathcal{P}, t)$ for the population $\mathcal{Y}$ is the expected loss defined by

$$R(\mathcal{P}, t; \mathcal{Y}) := E_{\mathcal{P}} L(t, \mathcal{Y}) = \sum_S \mathcal{P}(S) L(t(D[S, \mathcal{Y}_S]), \mathcal{Y}) \tag{3}$$

where the sum extends over all subsets of $\mathcal{N}$.

An important special case is $R(\mathcal{P}, t; \mathcal{Y}) := \mathrm{MSE}(\mathcal{P}, t; \mathcal{Y}) := E_{\mathcal{P}}(t - \theta)^2$; one reason for the interest in this measure is that by Chebychev's inequality it provides a lower bound on confidence interval coverage: $\mathcal{P}(|t - \theta(\mathcal{Y})| \leq c) \geq 1 - MSE(\mathcal{P}, t; \mathcal{Y})/c^2$ for each $\mathcal{Y} \in \Upsilon$. For unbiased estimators, the MSE coincides with the variance, which plays a role in the construction of confidence intervals based on the normal approximation. It is well known that the MSE of an estimator can be decomposed into the sum of its variance and the square of its bias.

8. The strategy $(\mathcal{P}, t)$ is said to be **unbiased** for $\theta = \theta(\mathcal{Y})$ if

$$E_{\mathcal{P}} t := \sum_S \mathcal{P}(S) t(D[S, \mathcal{Y}_S]) = \theta(\mathcal{Y}) \tag{4}$$

for all $\mathcal{Y} = (y_1, \ldots, y_N) \in \Upsilon$. In this case we say that t is $\mathcal{P}$-**unbiased**.

Note that if $\mathcal{P}$ satisfies $\alpha_i = n/N$ for all $i = 1, \ldots, N$, then the sample mean $\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i$ is unbiased for the population average $\bar{\mathcal{Y}}$ and, more generally for any design $\mathcal{P}$, so is the estimator t_{HT}/N where $t_{HT} = \sum_{i \in S} y_i/\alpha_i$ is the Horvitz-Thompson estimator, since by setting I_i to be the indicator of the event that $i \in S$ we have $E_{\mathcal{P}} I_i = \alpha_i$ and therefore

$$E_{\mathcal{P}}[t_{HT}/N] = E_{\mathcal{P}} \frac{1}{N} \sum_{i=1}^N I_i y_i/\alpha_i = \frac{1}{N} \sum_{i=1}^N y_i E_{\mathcal{P}} I_i/\alpha_i = \bar{\mathcal{Y}}. \tag{5}$$

More generally, the estimator $t = \frac{1}{N} \sum_{i \in S} y_i c_i(S)/\alpha_c(i)$, where $\alpha_c(i) = \sum_{S : S \ni i} c_i(S) \mathcal{P}(S)$,

is easily seen to be $\mathcal{P}$-unbiased for $\bar{\mathcal{Y}}$. When $c_i(S) \equiv 1$ it reduces to t_{HT}.

Under SRS, the ratio estimator $t_R = (\bar{y}_S/\bar{x}_S)\bar{\mathcal{X}}$ is, in general, not unbiased. On the other hand, the strategy $(\mathcal{P}_{ppas}, t_R)$, with $\mathcal{P}_{ppas}$ defined above as probability proportional to $\sum_{i \in S} x_i$ sampling, is unbiased for $\bar{\mathcal{Y}}$, since

$$E_{ppas}\, t_R = \sum_S (\sum_{i \in S} x_i) / \left[N\bar{\mathcal{X}} \binom{N-1}{n-1} \right] (\bar{y}_S/\bar{x}_S)\bar{\mathcal{X}} = \binom{N-1}{n-1}^{-1} \frac{1}{N} \sum_S \sum_{i \in S} y_i = \bar{\mathcal{Y}}.$$

To compare the above notions for finite populations with standard statistical decision theory, we give the following concise definitions, to be followed by a short discussion. For further details see, for example, Ferguson (1967) and Lehmann and Casella (1998).

Definition 3 • *An **observation** is a random variable $X \sim P_\theta$ (i.e., X has the distribution P_θ), where $\theta \in \Theta$, the **parameter space**.*

- *A **decision rule** $\delta(X)$ is a function taking values in a decision space $\mathcal{A}$, or a distribution on $\mathcal{A}$ (which may depend on X) in which case δ is **randomized**. The decision space is sometimes identical to the parameter space.*

- *$L(a, \theta)$ is the **loss** due to a decision $a \in \mathcal{A}$, and if δ is randomized, we set $L(\delta, \theta) = E_\delta L(a, \theta)$ where $a \sim \delta = \delta(X)$. The **risk** is defined by $R(\delta, \theta) = EL(\delta(X), \theta)$, where the expectation is with respect to $X \sim P_\theta$. Given a prior distribution ρ for θ, the **Bayes risk** is $r(\rho, \delta) = E_\rho R(\delta, \theta) = \int R(\delta, \theta) d\rho(\theta)$.*

- *A decision rule δ_0 is **Bayes with respect to** ρ if $r(\delta_0, \rho) = \inf_\delta r(\delta, \rho)$. It is a **minimax** rule if $\sup_\theta R(\delta_0, \theta) = \inf_\delta \sup_\theta R(\delta, \theta)$. The rule δ_0 has a **uniformly minimal risk** among **unbiased** estimators of $g(\theta)$ if $E\delta_0(X) = g(\theta)$, that is, δ_0 is unbiased, and $R(\delta_0, \theta) \leq R(\delta, \theta)$ for all $\theta \in \Theta$ and any unbiased rule δ. In the case of MSE risk, the latter δ_0 has the Uniformly Minimum Variance among Unbiased estimators (UMVU) property.*

In standard decision theory as given in Definition 3, the distribution of the data is prescribed as part of the problem, and optimization is done only with respect to the decision rule or the estimator. In contrast, when we study strategies in finite population sampling, we attempt to optimize over both the estimator and the sampling design. The latter determines the data collection method and the distribution of the data, and in this sense optimality in finite population sampling is more comprehensive than classical decision theory.

Remark 4 *Henceforth we consider only **nonrandomized** estimators unless otherwise stated (as when we consider non-convex loss in Section 3.4, and the Rao-Hartley-Cochran strategy in Section 7.2). When the loss function $L(a, \theta)$ (or $L(\tau, \mathcal{Y})$) is convex in the variable a (or τ), as in the quadratic loss case, then randomized estimators can be replaced by nonrandom ones having a smaller risk. In fact, by Jensen's inequality the risk of a randomized estimator can only decrease when the estimator is replaced by its expectation (assuming it is finite), which is a nonrandomized estimator.*

One may now ask whether randomization in the sampling design can also be eliminated in a similar way under some convexity conditions, that is, can a design $\mathcal{P}$ be replaced by a deterministic sample with a smaller risk. However, this cannot be done since the relevant space is not convex: there is no "average" or "expected" set for a given design.

3. Minimax strategies

3.1. Definitions and discussion

Definition 5 *A strategy $(\mathcal{P}_0, t_0)$ is said to be **minimax** relative to a given class of strategies if it belongs to this class, and*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_0, t_0; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{6}$$

for every strategy $(\mathcal{P}, t)$ in the given class of strategies.

*The estimator t_0 is said to be **minimax under** $\mathcal{P}$ in a class of estimators, if*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t_0; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{7}$$

for any estimator t in the class.

In (6) and below, the sup may be replaced by max when the latter exists. With quadratic loss function $\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y})$ becomes $\sup_{\mathcal{Y} \in \Upsilon} \text{MSE}(\mathcal{P}, t; \mathcal{Y})$, which is $\sup_{\mathcal{Y} \in \Upsilon} \text{Var}_{\mathcal{P}}(t)$ for unbiased estimators.

A minimax strategy guarantees the lowest maximal risk, that is, the smallest risk in the *worst case* or the worst possible $\mathcal{Y}$. If we denote the left-hand side of (6) by v_0, then using the strategy $(\mathcal{P}_0, t_0)$ we are guaranteed a risk of at most v_0 whatever the population values $\mathcal{Y}$ are, and a lower value *for all* $\mathcal{Y}$ cannot be guaranteed.

It turns out that minimax strategies involve random sampling. Strategies that avoid randomization are in general not minimax and hence may yield very poor estimates for certain populations $\mathcal{Y}$. Randomization guarantees that the sample "represents" the population (with high probability). Any fixed sample could be very biased relative to certain

populations. For example, the mean of a sample consisting of the first n labels from an ordered $\mathcal{Y}$ would be a poor estimate of the population mean, and such poor samples are avoided with high probability by randomization. This is why regulatory agencies insist on randomization, and perhaps also in order to prevent biased experimenters who have some partial knowledge of $\mathcal{Y}$ from choosing a biased sample that would prove their point rather than yield good estimates.

Minimax strategies are particularly relevant in zero-sum games, where maximizing one's own gain is equivalent to minimizing one's opponent's gain. The view of a statistical problem as a game between a statistician who chooses a strategy $(\mathcal{P}_0, t_0)$ and *nature* which "chooses" the parameter value, appears in well-known texts such as Blackwell and Girshick (1954) and Ferguson (1967). Random sampling is equivalent to a mixed strategy of the statistician, that is, a strategy which chooses the action (in this case, the sample S) at random according to a certain probability law (which in our case is $\mathcal{P}$). In general, minimax strategies are mixed strategies. Thus the minimax criterion leads naturally to random sampling. One may argue that nature should not be considered a strategic player who uses the worst possible (for the statistician) or least favorable $\mathcal{Y}$ as a player in a zero-sum game, and question the minimax approach and the relevance of zero-sum games. However, the protection against a worst-case population appears quite reasonable when prior knowledge of the populations is very limited.

Brewer (1963) and Royall (1970b) present optimality results where the sup's in (6) are replaced by expectations with respect to a prior (superpopulation model) on $\mathcal{Y}$ satisfying certain conditions that are expressed in terms of covariates. The resulting optimal design, which may be very sensitive to the choice of a prior, is nonrandom: averaging over a prior replaces the need for averaging by a random design. This approach is analogous to average-case or probabilistic analysis of algorithms in computer science, whereas the minimax approach pertains to worst-case evaluations.

While protecting against the worst case in the parameter space, minimax rules may sometimes be relatively unsatisfactory in other parts of that space. An example is given in Section 3.5.

3.2. Some minimax results through symmetry (invariance)

Invariance or symmetry has a long history in statistics. Symmetrization of strategies (as in (9) below) appears in Blackwell and Girshick (1954), and in Kiefer (1957) with reference to work of Hunt and Stein from the 1940s.

The first step towards finding minimax strategies through symmetry is to show that it suffices to search among strategies consisting of symmetric estimators and a (conditional) SRS design. This is formulated in Proposition 6. Part of the notation below follows

Stenger (1979).

Let Π denote the group of permutations of $\mathcal{N} = \{1, \ldots, N\}$, and for $\pi \in \Pi$, $S \subseteq \mathcal{N}$, and $\mathcal{Y} = (y_1, \ldots, y_N)$, define

$$\pi S = \{\pi(i) : i \in S\}, \quad \pi \mathcal{Y} = (y_{\pi^{-1}(1)}, \ldots, y_{\pi^{-1}(N)}). \tag{8}$$

For a design $\mathcal{P}$ let

$$\bar{\mathcal{P}}(S) = \sum_{\pi \in \Pi} \mathcal{P}(\pi S)/N!, \quad \bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = \frac{1}{N!\bar{\mathcal{P}}(S)} \sum_{\pi \in \Pi} t(\pi S, \pi \mathcal{Y})\mathcal{P}(\pi S). \tag{9}$$

Note that $\bar{\mathcal{P}}(S)$ is a probability on subsets S of $\mathcal{N}$. If the design $\mathcal{P}$ concentrates on sets of size n, then $\bar{\mathcal{P}}$ is a uniform probability with $\bar{\mathcal{P}}(S) = 1/\binom{N}{n}$ on such sets, that is, $\bar{\mathcal{P}}(S) = \mathcal{P}_s(S)$, which is SRS. If $\mathcal{P}(|S| = m) = \gamma_m$, $m = 1, \ldots, N$, where $|S|$ denotes the size of S, then $\bar{\mathcal{P}}(S) = \gamma_{|S|}/\binom{N}{|S|}$. In this case $\bar{\mathcal{P}}(S)$ is uniform over all sets of a given size, and we call it a ***conditional*** SRS. Moreover $\bar{\mathcal{P}} = \mathcal{P}$ if and only if $\mathcal{P}$ is a conditional SRS.

Let us now consider a random pair (π, S), consisting of a random permutation and a random set having the joint distribution $(\pi, S) \sim \frac{\mathcal{P}(\pi S)}{N!}$. It is easy to see that $\sum_\pi \sum_S \frac{\mathcal{P}(\pi S)}{N!} = 1$, and by (9) we have $\sum_\pi \frac{\mathcal{P}(\pi S)}{N!} = \bar{\mathcal{P}}(S)$, the marginal distribution of S. The conditional distribution of π given S is the ratio of the joint distribution $\frac{\mathcal{P}(\pi S)}{N!}$ and marginal distribution $\bar{\mathcal{P}}(S)$ of S. We summarize this notation as follows:

$$(\pi, S) \sim \frac{\mathcal{P}(\pi S)}{N!}, \qquad S \sim \bar{\mathcal{P}}(S), \qquad \pi | S \sim \frac{\mathcal{P}(\pi S)}{N!\bar{\mathcal{P}}(S)}. \tag{10}$$

Then $\bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = E[t(\pi S, \pi \mathcal{Y}) \,|\, S] = E_{\pi|S}[t(\pi S, \pi \mathcal{Y})]$.

Note that $D[\pi S, \pi \mathcal{Y}] = \{(\pi(i), y_{\pi^{-1}\pi(i)}) : i \in S\} = \{(\pi(i), y_i) : i \in S\}$, and so $\bar{t}_{\mathcal{P}}(S, \mathcal{Y})$ does not depend on y-values outside of $\mathcal{Y}_S$. The same is true for any $t(\pi S, \pi \mathcal{Y})$ with known π. Assume without loss of generality that $S = \{1, \ldots, n\}$, and use the notation $\pi(i) = j_i$ for $i \in S$. From (9) we have

$$\bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = c \sum_{\{j_1, \ldots, j_n\}} t(\{(j_1, y_1), \ldots, (j_n, y_n)\})\mathcal{P}(\{j_1, \ldots, j_n\}),$$

where c is a constant. The sum is over all subsets of size n and does not depend on S, and it follows that $\bar{t}_{\mathcal{P}}(S, \mathcal{Y})$ depends on $\mathcal{Y}_S$ (and $\mathcal{P}$), but not on S; that is, $\bar{t}_{\mathcal{P}}$ is a symmetric estimator.

It is now easy to see that for a symmetric estimator $t(\mathcal{Y}_S)$ we have

$$t(\pi S, \pi \mathcal{Y}) = t(S, \mathcal{Y}) \quad \text{and therefore} \quad \bar{t}_\mathcal{P}(S, \mathcal{Y}) = t_\mathcal{P}(S, \mathcal{Y}). \tag{11}$$

From the definitions in (9) it is easy to see that if $(\mathcal{P}, t)$ is an **unbiased strategy** for a *symmetric* parameter $\theta = \theta(\mathcal{Y})$, that is, a parameter satisfying $\theta(\pi \mathcal{Y}) = \theta(\mathcal{Y})$ for all $\pi \in \Pi$, then so is the strategy $(\bar{\mathcal{P}}, \bar{t})$.

The following proposition (see Gabler (1990) and references therein for closely related results) implies that for a minimax strategy relative to designs of fixed sample size, it suffices to search among strategies $(\mathcal{P}_s, t)$, where $t = t(\mathcal{Y}_S)$ is symmetric and $\mathcal{P}_s$ is SRS with the same sample size. More generally, it shows that for any strategy $(\mathcal{P}, t)$ there is a strategy consisting of a conditional SRS design having the same distribution of sample size as $\mathcal{P}$ and a symmetric estimator $t(\mathcal{Y}_S)$ with a smaller maximal risk. The proposition requires symmetry and convexity of L. A closely related result that does not require convexity of the loss is Proposition 13. The latter proposition provides an interpretation of the third expression in (13) below in terms of a randomized estimator. We defer the discussion to Section 3.4 for the sake of simplicity at this point.

Recall that a loss function L is symmetric if it remains constant under permutations of $\mathcal{Y}$; that is, $L(\tau, \mathcal{Y}) = L(\tau, \pi \mathcal{Y})$. With the above definitions we have

Proposition 6 *Let $L(\tau, \mathcal{Y})$ be a symmetric loss function that is convex in τ for each $\mathcal{Y} \in \Upsilon$, a symmetric parameter space. Then for $\bar{\mathcal{P}}, \bar{t}$ defined in (9)*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}). \tag{12}$$

Proof

$$
\begin{aligned}
R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) &= \sum_S L(\bar{t}(S, \mathcal{Y}), \mathcal{Y}) \bar{\mathcal{P}}(S) \overset{(0)}{\leq} \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N! \bar{\mathcal{P}}(S)} L(t(\pi S, \pi \mathcal{Y}), \mathcal{Y}) \bar{\mathcal{P}}(S) \\
&= \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!} L(t(\pi S, \pi \mathcal{Y}), \mathcal{Y}) \overset{(1)}{=} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi \mathcal{Y}), \mathcal{Y}) \\
&\overset{(2)}{=} \frac{1}{N!} \sum_\pi \sum_S \mathcal{P}(S) L(t(S, \pi \mathcal{Y}), \pi \mathcal{Y}) = \frac{1}{N!} \sum_\pi R(\mathcal{P}, t; \pi \mathcal{Y}); \tag{13}
\end{aligned}
$$

Jensen's inequality applied to the convexity of the loss function L implies the inequality marked by (0); a further explanation is given below. The equality (1) was obtained by substituting S for πS (both range over all subsets of $\mathcal{N}$ under the summation on S) and (2) follows because by symmetry $L(\tau, \mathcal{Y}) = L(\tau, \pi \mathcal{Y})$.

A simple way to understand the above inequality (0) is to note that under the definitions

of (10),

$$\sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!\bar{\mathcal{P}}(S)} L(t(\pi S, \pi \mathcal{Y}), \mathcal{Y})\bar{\mathcal{P}}(S) = E_S\{E_{\pi|S}[L(t(\pi S, \pi \mathcal{Y}), \mathcal{Y})]\},$$

and the inequality becomes $E_S L(E_{\pi|S}[t(\pi S, \pi \mathcal{Y})], \mathcal{Y}) \le E_S\{E_{\pi|S}[L(t(\pi S, \pi \mathcal{Y}), \mathcal{Y})]\}.$

From (13) we have $R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) \le \max_\pi R(\mathcal{P}, t; \pi \mathcal{Y})$ since the maximum is larger than the average, and by the symmetry (permutation invariance) of the parameter space Υ, it follows that

$$\sup_{\mathcal{Y} \in \Upsilon} R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) \le \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}). \qquad \blacksquare$$

Note that for a (conditional) SRS design $\mathcal{P}$ we have $\mathcal{P} = \bar{\mathcal{P}}$, and we conclude from Proposition 6 that for such designs it suffices to consider symmetric estimators when the goal is to minimize maximal risk with convex loss. This is formulated in the corollary below, which appears in Royall (1970a).

Corollary 7 *Let $\mathcal{P}$ be a conditional SRS design. Under the conditions of Proposition 6*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, \bar{t}; \mathcal{Y}) \le \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}).$$

Our next goal is to establish a minimax result for unbiased strategies. First we need the following lemma on completeness, due to Royall (1968). As usual, completeness will be used to obtain uniqueness of unbiased estimators. A function $h(y_1, \dots, y_n)$ is said to be **symmetric** if it is invariant under permutations of its arguments, that is, it depends only on the set of (unordered) values $\{y_1, \dots, y_n\}$. We can then write $h(\mathcal{Y}_S)$, since $\mathcal{Y}_S$ is a set of (unordered) y-values.

Lemma 8 *Let Υ be any product parameter space*

$$\Upsilon = \Lambda^N = \Lambda \times \cdots \times \Lambda, \tag{14}$$

and let $\mathcal{P}$ be a design such that $\mathcal{P}(S) > 0$ implies $|S| = n$. Then $\mathcal{Y}_S$ is complete; that is, for any symmetric function h, $E_\mathcal{P} h(\mathcal{Y}_S) = 0$ for all $\mathcal{Y} = (y_1, \dots, y_N) \in \Upsilon$ implies that $h(y_1, \dots, y_n) = 0$ for any $y_i \in \Lambda$, $i = 1, \dots, n$.

Proof First consider $\mathcal{Y} = (a, \dots, a) \in \Upsilon$ (here, for example, we use the structure of Υ given by (14)) and compute the expectation under this value of the parameter. Then $0 = E_\mathcal{P} h(\mathcal{Y}_S) = \sum_S \mathcal{P}(S)h(a, \dots, a)$ implies $h(a, \dots, a) = 0$. Assuming without loss of generality that $\mathcal{P}(S) > 0$ for $S = \{1, \dots, n\}$ choose now $\mathcal{Y} = (b, a, \dots, a) \in \Upsilon$.

Then $0 = E_{\mathcal{P}} h(\mathcal{Y}_S) = ph(a, \ldots, a) + qh(b, a, \ldots, a)$ with $q > 0$, and we conclude that $h(b, a, \ldots, a) = 0$. The result follows by continuing in the same manner (induction). ∎

The next theorem shows that relative to the class of unbiased strategies, there exist minimax strategies that involve simple random sampling without replacement (SRS). For a closely related result see Theorem 3.10 in Cassel, Särndal, and Wretman (1977) and references therein. Remark 11 compares their result to Theorem 9 below. Such a result is not true without restricting the class to unbiased strategies (see Remark 11 below). Unbiasedness is ubiquitous in applications. This is quite natural since avoiding bias is often given as a justification for random sampling. However, unbiasedness alone does not guarantee good estimation; see, e.g., Basu's (1971) famous circus-elephants weighing example for a ridiculously poor unbiased estimator.

Theorem 9 *Let Υ be any product parameter space: $\Upsilon = \Lambda^N = \Lambda \times \cdots \times \Lambda$, and let $\mathcal{P}_s$ denote simple random sampling without replacement (SRS) of size n. If there exists any unbiased strategy $(\mathcal{P}, t)$ for the parameter $\theta = \theta(\mathcal{Y})$ with $\mathcal{P}$ having a fixed sample size n, then there exists a unique symmetric estimator $t_0 = t_0(\mathcal{Y}_S)$ depending only on $\mathcal{Y}_S$, such that the strategy $(\mathcal{P}_s, t_0)$ is unbiased.*

If $\theta = \theta(\mathcal{Y})$ is a symmetric parameter, and the loss function $L(\tau, \theta)$ is convex in τ for each $\mathcal{Y} \in \Upsilon$, then

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, t_0; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{15}$$

*for any unbiased strategy $(\mathcal{P}, t)$ for θ, having a fixed samples size n. In other words, the strategy $(\mathcal{P}_s, t_0)$ is **minimax** relative to the above class of strategies $(\mathcal{P}, t)$.*

Proof As mentioned after (11), if $(\mathcal{P}, t)$ is unbiased then so is $(\bar{\mathcal{P}}, \bar{t})$. Since $\mathcal{P}$ concentrates on sets of size n, we have $\bar{\mathcal{P}} = \mathcal{P}_s$. Lemma 8 implies that a symmetric unbiased estimator is unique (take h in the lemma to be the difference between two unbiased estimates to obtain that they are the same). It follows that the strategy $(\bar{\mathcal{P}}, \bar{t})$ is the same for all unbiased $(\mathcal{P}, t)$, and the result follows from (12) with $t_0 = \bar{t}$.

The sup's in (15) may be infinite, in which case the result is uninteresting, and it is empty if no unbiased strategies exist. ∎

Corollary 10 *Let $\Upsilon = \Lambda^N$. If $\theta = \theta(\mathcal{Y}) = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{\mathcal{Y}}$, the population mean, and the loss function $L(\tau, \theta)$ is convex in τ for each $\mathcal{Y} \in \Upsilon$, then for the sample mean $\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i$ we have*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{16}$$

*for any unbiased strategy $(\mathcal{P}, t)$ for θ, having a fixed sample size n. In other words, the strategy $(SRS, \bar{y}_S)$ is **minimax** relative to the above class of strategies $(\mathcal{P}, t)$.*

For the population variance $\theta = V(\mathcal{Y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{\mathcal{Y}})^2$, set $v = \frac{N-1}{N(n-1)} \sum_{i \in S} (y_i - \bar{y}_S)^2$, an unbiased estimator. Then (SRS, v) is minimax relative to unbiased strategies having sample size n.

Proof The population mean $\theta = \bar{\mathcal{Y}}$ is a symmetric parameter, and the sample mean $\bar{y}_S$ is a $\mathcal{P}_s$-unbiased estimator, that is, it is unbiased for SRS of size n. The result follows from Theorem 9. Similarly, v above is symmetric and a $\mathcal{P}_s$-unbiased estimator of the population variance, which is a symmetric parameter. ■

Remark 11 *Theorem 3.10 of Cassel, Särndal, and Wretman (1977) states a result similar to Theorem 9 for the special case of $\theta = \bar{\mathcal{Y}}$, the population mean, and for the quadratic loss function (MSE). It states that in this case an unbiased strategy $(\mathcal{P}_1, \bar{y}_S)$ with sample size n is minimax relative to the class of unbiased strategies with sample size n, for any such $\mathcal{P}_1$ satisfying $\alpha_i = n/N$ for $i = 1, 2, \ldots, N$, and it seems that all they require of the parameter space is for it to be symmetric.*

In the counterexamples below we also consider quadratic loss and estimation of the population mean. The first example shows that the assumption $\alpha_i = n/N$ does not suffice.

Set $N = 4$, $n = 2$, and $\Upsilon = \{0, 2a\}^4$. Then for quadratic loss a straightforward calculation shows that $\max_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) = a^2/2$ and the maximum is attained at $\mathcal{Y} = (0, 0, 0, 2a)$. On the other hand the design defined by $\mathcal{P}_1(\{1, 2\}) = \mathcal{P}_1(\{3, 4\}) = 1/2$ satisfies $\alpha_i = n/N = 1/2$, but for $\mathcal{Y} = (0, 0, 2a, 2a)$ one easily gets $R(\mathcal{P}_1, \bar{y}_S; \mathcal{Y}) = a^2$ and clearly $(\mathcal{P}_1, \bar{y}_S)$ is not minimax.

The next example shows that even in the case of SRS it is not enough to assume that Υ is symmetric. Set $\Upsilon = \Omega(1, 2, 3) \bigcup \Omega(11, 12, 13)$, that is, the set consisting of the two indicated vectors and all their permutations. Here $N = 3$. Then for SRS with $n = 1$ or $n = 2$, there clearly exists an (unbiased) estimator t which is always exactly correct, and hence satisfies $R(\mathcal{P}_s, t; \mathcal{Y}) = 0$ whereas $R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) > 0$, so that $(\mathcal{P}_s, \bar{y}_S)$ is not minimax. This happens because one observation provides complete information about the population up to permutations (recall that the parameter space is assumed known).

*Finally, we show by a simple example that the unbiasedness condition is not redundant. (For MSE, this will become clear also in Section 3.5.) In fact, a **biased** estimate t may satisfy $\max_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) < \max_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y})$. Take $N = 2, \Upsilon = \{0, 1\}^2$, and $n = 1$. Then $\max_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) = (1/2)^2$. The (biased) estimator t defined by $t(0) = 1/4$, $t(1) = 3/4$, satisfies $\max_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) = (1/4)^2$.*

The restriction to unbiased estimators may be replaced by linearity and invariance conditions and similar results to Theorem 9 still hold. Note that for the case of estimating $\bar{\mathcal{Y}}$,

for example, the minimax strategy $(SRS, \bar{y}_S)$ *does not depend on* Υ. *Without unbiasedness or similar restrictions, the minimax strategy depends on* Υ, *and finding it may be difficult. For nonsymmetric parameter spaces the problem becomes even harder. See Proposition 17 below for a minimax rule on symmetric product parameter spaces for quadratic loss (MSE), without an unbiasedness condition.*

3.3. Symmetric estimators and non-convex loss

The next proposition is a special case of a general result on invariance, Theorem 8.6.4 of Blackwell and Girshick (1954), who applied it in the context of sampling. It says that for **symmetric estimators** the maximal risk is minimized by (conditional) SRS designs. In particular, designs having a fixed sample size can be replaced by SRS. Since $R(\mathcal{P}, t; \mathcal{Y})$ is linear in $\mathcal{P}$ (but not in t), **convexity of the loss L is not required**. Also, we require no conditions on Υ other than symmetry, which is always assumed. Recall that for a given design $\mathcal{P}$ the corresponding $\bar{\mathcal{P}}$ is defined in (9).

Proposition 12 *For any symmetric estimator t, design $\mathcal{P}$, and symmetric loss function L,*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\bar{\mathcal{P}}, t; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}).$$

If the design $\mathcal{P}$ has a fixed sample size, say n, then

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}),$$

where $\mathcal{P}_s$ denotes SRS of size n.

Proof

$$
\begin{aligned}
\sup_{\mathcal{Y} \in \Upsilon} R(\bar{\mathcal{P}}, t; \mathcal{Y}) &= \sup_{\mathcal{Y} \in \Upsilon} \sum_S \bar{\mathcal{P}}(S) L(t(S, \mathcal{Y}), \mathcal{Y}) = \sup_{\mathcal{Y} \in \Upsilon} \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!} L(t(S, \mathcal{Y}), \mathcal{Y}) \\
&= \sup_{\mathcal{Y} \in \Upsilon} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(\pi^{-1}S, \mathcal{Y}), \mathcal{Y}) \overset{(1)}{=} \sup_{\mathcal{Y} \in \Upsilon} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi\mathcal{Y}), \mathcal{Y}) \\
&\overset{(2)}{=} \sup_{\mathcal{Y} \in \Upsilon} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi\mathcal{Y}), \pi\mathcal{Y}) \leq \frac{1}{N!} \sum_\pi \sup_{\mathcal{Y} \in \Upsilon} \sum_S \mathcal{P}(S) L(t(S, \pi\mathcal{Y}), \pi\mathcal{Y}) \\
&= \frac{1}{N!} \sum_\pi \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}) = \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}); \quad (17)
\end{aligned}
$$

the equalities marked by (1) is obtained by the symmetry of t using the first part of (11), and (2) follows from the symmetry of L; the rest is straightforward. The second part of the proposition is a special case of the first, based on the fact that if $\mathcal{P}$ has a fixed sample size n, then $\bar{\mathcal{P}} = \mathcal{P}_s$. ∎

It is easy to provide a counterexample to the above result for t that is not symmetric. Take $N = 3$ and $n = 2$ and let $t(S, \mathcal{Y})$ be an estimator that takes a huge and irrelevant value when $3 \in S$. Clearly one can choose values such that the design satisfying $\mathcal{P}(\{1, 2\}) = 1$ will violate the second inequality of Proposition 12.

3.4. Asymmetric and randomized estimators and non-convex loss

So far we have considered loss functions $L(\tau, \mathcal{Y})$ that are **convex** in τ, with one exception, namely Proposition 12. In sample survey applications, unbiased or nearly unbiased estimators are usually considered, and their variances or MSE are computed. This corresponds to quadratic loss, which is convex. In this section we shall see (and it is well known) that for non-convex loss functions and certain optimality criteria of statistical decision theory, randomized estimators become relevant, and we discuss them briefly in the next paragraph. Non-convex loss functions arise, for example, in specific areas such as statistical classification, where a convex loss would tend to overemphasize misclassification of outliers, and more generally, when one wants to allow bounded loss over unbounded spaces. In this chapter we treat general loss functions, including non-convex ones, because we think that they may be useful and relevant, and because their discussion clarifies the analysis and shows what conditions are really needed, an issue that may be hidden in explicit calculations with quadratic loss.

For randomized estimators, standard decision theory suggests taking expectation of the loss over both the random estimator, and the design. This leads to the interpretation of the loss for randomized estimators as given in Definition 3: $L(\delta, \theta) = E_\delta L(a, \theta)$ where $a \sim \delta = \delta(X)$. This interpretation, which in certain situations leads to optimality of randomized estimators as shown later, is perhaps relevant when a large number of similar estimation problems are considered together, with (roughly) the same value of the estimated parameter. The law of large numbers is then often used to justify the expectation. Perhaps one may consider repeated estimation of employment rates in a monthly Labor Force Survey to be such a situation. But in general, statistical agencies do not use randomized estimators, and their discussion in our context is theoretical.[3] The latter fact indicates that this approach to loss and randomized estimators is debatable. For example, when the randomization does not depend on the data, which is the case in most of the examples given later, this interpretation seems to violate the conditionality principle which many statisticians accept[4], since it takes into account possible estimators which were not chosen to be used.

---

[3] However, the Rao-Hartley-Cochran strategy mentioned in Section 7.2 is an example of a randomized estimator.

[4] See Helland (1995) for the history, a critical discussion, and references on the conditionality principle.

Like random sampling (which is, of course, used everywhere) randomized estimators may be seen as mixed strategies in game theoretical terminology; see, e.g., Rubinstein (1991) and references therein for a discussion of the difficulty of interpreting mixed strategies in game theory, which pertains to statistics as well.

Below is a simple example showing that without convexity, randomized estimators must sometimes be taken into account. Consider for example the loss function of a *perfectionist* defined by $L(\tau, \theta) = 0$ if $\tau = \theta$ and $= 1$ otherwise,[5] and let θ be the population mean. Let $n = 1$, $\Upsilon = \{0, 1\}^2$, that is, $N = 2$. Then under SRS, for example, the randomized estimator t^* with $t^*(0) = 0$ or $= 1/2$ with probability $1/2$ each, and $t^*(1) = 1/2$ or $= 1$, again with probability $1/2$, is the minimax rule with risk $= 1/2$. For convex loss function, a simple application of Jensen's inequality implies that we would achieve the same or smaller risk by averaging t^* to obtain the estimator t with $t(0) = 1/4, t(1) = 3/4$ (see Remark 11), which for quadratic loss is minimax. However, for the perfectionist's loss function the risk of t equals 1; it is an estimator that is never exactly correct.

The next result says that for any symmetric loss function (convex or not) and any strategy $(\mathcal{P}, t)$ having a fixed sample size n, one can find an estimator t^* such that the maximal risk of $(\mathcal{P}_s, t^*)$ is smaller than that of $(\mathcal{P}, t)$, where $\mathcal{P}_s$ denotes SRS of size n. This suggests that for minimax purposes or when considering maximal risk (and with the absence of auxiliary information), only SRS needs to be considered.

The estimator t^* turns out to be randomized, and its construction is given explicitly in Proposition 13 below. I cannot provide a reference for this proposition; it is probably not new, but if it is, it may be because little or no attention has been paid to non-convex loss functions in finite population sampling. See Ferguson (1967 Theorem 4.3.1) for a related result, where randomized rules play a similar role. The proposition shows that the fact that SRS suffices for minimax considerations when estimating a symmetric parameter (which implies symmetric loss) is not related to convexity.

Given an estimator t, let $t_\pi(S, \mathcal{Y}) = t(\{(\pi(i), y_i) : i \in S\} = t(\pi S, \pi \mathcal{Y})$; see (1) and (8) for notations. For example if $t = t_{HT}$ then $t_\pi(S, \mathcal{Y}) = \sum_{i \in S} y_i / \alpha_{\pi(i)}$. For a strategy $(\mathcal{P}, t)$ with a fixed sample size, let t^* be the **randomized** estimator defined for a given S by

$$t^*(S, \mathcal{Y}) = t_\pi(S, \mathcal{Y}) = t(\pi S, \pi \mathcal{Y}) \quad \text{with probability} \quad \frac{\mathcal{P}(\pi S)}{N! \mathcal{P}_s(S)} \quad \text{for } \pi \in \Pi, \qquad (18)$$

where Π is the permutation group over $\{1, \ldots, N\}$. Note that the probabilities determining the randomization depend on S.[6] Clearly $\sum_\pi \frac{\mathcal{P}(\pi S)}{N! \mathcal{P}_s(S)} = 1$.

---

[5]A smoothed version of this function could be studied in similar ways.
[6]In this case, such a t^* is called a *behavioral estimator*.

Proposition 13 *Let $L(\tau, \mathcal{Y})$ be a symmetric loss function (convex or not) and as always let Υ be a symmetric parameter space. Given a strategy $(\mathcal{P}, t)$ with fixed sample size n, let $t^*(S, \mathcal{Y})$ be the* **randomized** *estimator of (18). Then*

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, t^*; \mathcal{Y}) \leq \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}). \tag{19}$$

Proof We just repeat part of the proof of Proposition 6. Using the notation of (8), but see also (9) and (13), we have

$$
\begin{aligned}
R(\mathcal{P}_s, t^*; \mathcal{Y}) &= \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N! \mathcal{P}_s(S)} L(t_\pi(S, \mathcal{Y}), \mathcal{Y}) \mathcal{P}_s(S) = \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!} L(t(\pi S, \pi \mathcal{Y}), \mathcal{Y}) \\
&\overset{(1)}{=} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi \mathcal{Y}), \mathcal{Y}) \overset{(2)}{=} \frac{1}{N!} \sum_\pi \sum_S \mathcal{P}(S) L(t(S, \pi \mathcal{Y}), \pi \mathcal{Y}) \\
&= \frac{1}{N!} \sum_\pi R(\mathcal{P}, t; \pi \mathcal{Y}) \leq \max_\pi R(\mathcal{P}, t; \pi \mathcal{Y}),
\end{aligned} \tag{20}
$$

where the relations marked by (1) and (2) are explained under (13). The result now follows easily (compare to the proof of Proposition 6). ∎

A similar result to the above holds when $\mathcal{P}$ does not have a fixed sample size, in which case the left-hand side of (19) holds with $\mathcal{P}_s$ replaced by the conditional SRS design $\bar{\mathcal{P}}$.

The relation between Propositions 13 and 6 is as follows. If the loss is convex we can replace t^* in (19) by its expectation, and obtain a lower bound by Jensen's inequality, and Proposition 6 follows; this is true also when the sample size is random.

We stated Propositions 12 and 13 separately only because in the context of finite population sampling, randomized estimators (which are needed to state Proposition 13) are esoteric. We could have stated just Proposition 13, since it implies Proposition 12 readily. To see this it suffices to note that by (11), if the estimator t is symmetric, then the estimator t^* defined in (18) is in fact nonrandomized, and $t^* = t$.

3.5. Minimax and Bayes estimators

Minimax estimators can be obtained from Bayesian calculations. An example of this approach concerning estimation of a proportion in a finite population is given with the purpose of demonstrating the technique. While minimizing the maximal risk by definition, the resulting minimax rule has a higher risk than the usual estimator, the sample proportion, in parts of the parameter space, and we discuss the comparison between the two estimators. Most of the following discussion and much more can be found in Lehmann and Casella (1998) and the references therein.

The problem of estimating a proportion in a finite population of size N by a sample of

size n is first approximated by the standard decision-theoretic problem of estimating the parameter p from a binomial distribution, that is, a sample of iid Bernoulli(p) observations. The notation and terminology we need for the latter problem is that of Definition 3.

3.5.1. The binomial case

Consider $X \sim \text{Binomial}(n, p)$ and a Bayesian structure with a prior $p \sim \text{Beta}(a, b)$. For quadratic loss, it is well known that the Bayes estimator is the posterior expectation $d(X) = E(p|X)$. A standard calculation shows that the estimator

$$d(X) = \frac{X}{n} \frac{\sqrt{n}}{1 + \sqrt{n}} + \frac{1}{2(1 + \sqrt{n})} \tag{21}$$

is Bayes with respect to the above prior when $a = b = \sqrt{n}/2$ and that it is an *equalizer*, that is, its risk is constant and does not depend on p. In fact, $E(d(X) - p)^2 = \frac{1}{4(1+\sqrt{n})^2}$. The following proposition is well known (see, e.g., Ferguson (1967) or Lehmann and Casella (1998)) and readily implies the minimax result of Corollary 15 below. For definitions see Definition 3.

Proposition 14 *A Bayes estimator δ_0 having a constant risk (equalizer) is minimax. If δ_0 is uniquely Bayes with respect to a given prior, then it is the unique minimax estimator.*

Proof Let δ be another estimator, and assume δ_0 is Bayes with respect to ρ. The estimator δ_0 satisfies $r(\delta_0, \rho) = \int R(\delta_0, \theta) d\rho \leq r(\delta, \rho) = \int R(\delta, \theta) d\rho$. As $R(\delta_0, \theta)$ is a constant not depending on θ, it follows that $R(\delta_0, \theta) \leq \int R(\delta, \theta) d\rho \leq \sup_\theta R(\delta, \theta)$ for all θ, and δ_0 is minimax. If another rule is minimax, then using the assumption of constant risk of δ_0, it is easy to see that it is also Bayes, and the uniqueness part follows. ∎

Corollary 15 *The estimator $d(X)$ of (21) is the unique minimax estimator of p for quadratic loss.*

For the estimator $d^*(X) = X/n$, which is UMVU, we have $E(d^*(X) - p)^2 = p(1-p)/n$, and we see that around $p = 1/2$ the estimator d is slightly better than d^* provided that n is not small, but d^* has smaller risk when p is not close to $1/2$. Thus here, and in the developments below where a similar phenomenon occurs, one may argue about the quality of the estimator obtained by the minimax criterion.

3.5.2. Finite population sampling for proportion and mean

We now follow Lehmann and Casella (1998) and Hodges and Lehmann (1982). Related results appear in Bickel and Lehmann (1981). In Corollary 16 and Proposition 17 below we obtain minimax results without restriction to unbiased estimators (compare to Corollary 10).

Consider SRS from a population of size N whose values are either 0 or 1, and we wish to estimate the parameter W/N where W is the number of ones. In fact, in this case $W/N = \bar{\mathcal{Y}}$. Consider the prior on W, which is a mixture of binomials with Beta(a, b) weights, that is,

$$P(W = w) = \int_0^1 \binom{N}{w} p^w (1-p)^{N-w} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp. \tag{22}$$

A reader who wants to avoid the calculations can look at numbered equations only. Let X be the number of ones drawn in an SRS of size n. Then $X|W \sim$ Hypergeometric, that is, $P(X = x|W) = \binom{W}{x}\binom{N-W}{n-x}/\binom{N}{n}$ and with standard calculations we have for some $c = c(x)$

$$
\begin{aligned}
P(W = w|X = x) &= cP(X = x|W = w)P(W = w) = c\int_0^1 \binom{N-n}{w-x} p^{w+a-1} q^{N-w+b-1} dp \\
&= c\int_0^1 \binom{N-n}{k} p^k q^{N-n-k} \cdot p^{x+a-1} q^{n-x+b-1} dp \, ;
\end{aligned}
$$

the last expression above is obtained by the substitution $k = w-x$ where $k = 0, \ldots, N-n$, and it is arranged so that under the integral we observe the Bin$(N-n, p)$ probability function in the variable k. It is now easy to see that $c = \Gamma(n+a+b)/[\Gamma(x+a)\Gamma(n-x+b)]$ is the normalizing constant so that $P(W = w|X = x)$ is a probability function. Using the Bin$(N-n, p)$ expectation we get

$$E(W - x|X = x) = c\int_0^1 (N-n)p \cdot p^{x+a-1} q^{n-x+b-1} dp = \frac{(N-n)(x+a)}{n+a+b},$$

and therefore we obtain that the Bayes estimator is the linear estimator

$$d(x) = E(W/N|X = x) = \frac{(N+a+b)x + (N-n)a}{N(n+a+b)}. \tag{23}$$

To compute the MSE of d, we use the relations $E(X|W) = nW/N$ and $\text{Var}(X|W) = Wn(N-W)(N-n)N^{-2}(N-1)^{-1}$ for the hypergeometric distribution of $X|W$, and the formula MSE(d)=Variance(d)+[Bias$(d)]^2$. We then choose a, b which make the MSE constant (not dependent on W). The resulting equalizer estimator is given in (24) below and we obtain

Corollary 16 *Under SRS with sample size n and quadratic loss, the estimator*

$$d(X) = AX/n + B, \quad where \ \ A = 1/[1 + \sqrt{(N-n)/(nN-n)} \,], \ \ B = (1-A)/2 \tag{24}$$

is minimax among symmetric estimators (depending only on X, the number of ones in the sample) for the proportion W/N of ones in a finite population of zeros and ones.

We omit the calculations; clearly the obtained estimator is minimax, being an equalizer and Bayes. Naturally the estimator obtained in (24) converges to that of (21) for large N, and has similar properties: it is worse than the usual sample mean when W is not near $N/2$, and it is somewhat better near $N/2$.

As already mentioned, for parameter spaces that are not symmetric, minimax estimators depend on the parameter space, and their calculation may be difficult in the absence of further restrictions on the decision rules. Corollary 16 provides a minimax rule for estimating a proportion, in which case the parameter space is $\{0,1\}^N$. We show next that for SRS on $\Upsilon = \Lambda^N$ and quadratic loss, Corollary 16 can be extended to provide a minimax estimator for any interval $\Lambda \subset \mathbb{R}$ (and more general sets). As noted by Lehmann and Casella (1998) (see references therein), this can be obtained from the previous discussion. See also Gabler (1990) for generalizations.

Proposition 17 *Let $\Upsilon = \Lambda^N$, where $\Lambda = [a, b]$ for some $a \leq b$. Let $\bar{\mathcal{Y}}$ and $\bar{y}_S$ denote the population and sample means. Under $\mathcal{P}_s = SRS$ with sample size n, and quadratic loss, the estimator $d_0 = (b - a)d((\bar{y}_S - a/(b - a)) + a$, where $d(z) = Az + B$ with A and B as defined in (24), is minimax for $\bar{\mathcal{Y}}$ relative to the class of all estimators. In the case $[a, b] = [0, 1]$ the minimax estimator is $d_0 = d(\mathcal{Y}_S) = A\bar{y}_S + B$. Moreover, the strategy $(\mathcal{P}_s, d_0)$ is minimax (see Definition 5) relative to the class of all strategies with a fixed sample size n.*

Proof We can assume first that $\Lambda = [0, 1]$, and then apply a linear transformation. By Corollary 7 we can restrict our attention to symmetric estimators t. By Corollary 16, the estimator of (24) is minimax among symmetric estimators when the parameter space is restricted to the set of extreme points of Υ, which we denote by $\Upsilon e = \{0,1\}^N$. Let E below denote expectation with respect to the (prior) probability measure on Υe defined by $P(\mathcal{Y} = (y_1, \ldots, y_N)) = P(W = w)/\binom{N}{w}$ for any vector $(y_1, \ldots, y_N) \in \Upsilon e$, where $\sum_{i=1}^N y_i = w$, $w = 0, 1, \ldots N$, and the distribution of W is given in (22). Note that the estimator $d = A\bar{y}_S + B$ is Bayes with respect to this prior and an equalizer on Υe. We have

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) \geq \sup_{\mathcal{Y} \in \Upsilon e} R(\mathcal{P}_s, t; \mathcal{Y}) \overset{(1)}{\geq} ER(\mathcal{P}_s, t; \mathcal{Y}) \overset{(2)}{\geq} ER(\mathcal{P}_s, d_0; \mathcal{Y})$$
$$\overset{(3)}{=} \sup_{\mathcal{Y} \in \Upsilon e} R(\mathcal{P}_s, d_0; \mathcal{Y}) \overset{(4)}{=} \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_s, d_0; \mathcal{Y});$$

inequality (1) holds because an average is smaller than the maximum, (2) holds because

d_0 is Bayes with respect to the given prior, and (3) holds because d_0 is an equalizer, that is, $R(\mathcal{P}_s, d; \mathcal{Y})$ is constant on Υe. Finally (4) follows from the fact that $R(\mathcal{P}_s, d; \mathcal{Y}) = \sum_S \mathcal{P}_s(S)(d(\mathcal{Y}_S) - \bar{\mathcal{Y}})^2$ is a convex function of $\mathcal{Y}$ and therefore its maximum is attained at the set of extreme points.

Proposition 6 readily implies that the strategy $(\mathcal{P}_s, d_0)$ is minimax as stated. ∎

It is easy to see that the above result holds for any bounded $\Lambda \subset \mathbb{R}$ satisfying $\{a, b\} \subseteq \Lambda \subseteq [a, b]$ for some $a, b \in \mathbb{R}$. By continuity arguments it also holds when $\Lambda = (a, b)$. If Λ is not convex, the estimator may take a value that is not in the parameter space, which is allowed, as our decision space is always $\mathbb{R}$. Proposition 17 is trivial if Λ is unbounded since the maximal risk is always infinite.

For the sample space $\Upsilon = \{\mathcal{Y} : \sum_{i=1}^N (y_i - \bar{\mathcal{Y}})^2 \leq M\}$, Bickel and Lehmann (1981) proved that under simple random sampling, the sample mean is minimax for quadratic loss. Since the variance of the sample mean is proportional to $\sum_{i=1}^N (y_i - \bar{\mathcal{Y}})^2$, the definition of the sample space is equivalent to assuming that this variance is bounded by a given constant. The proof uses invariance, and a reduction of the problem to estimation of a translation parameter, and showing that the sample mean coincides with the Pitman estimator. If the population is divided into given **strata**, then the usual weighted average of the sample means in the strata is minimax when the parameter space is defined by the condition that its variance is bounded by a given constant. Results on optimal designs are also given.

Related results by Aggarwal (1959, 1966) for a superpopulation model are described in Section 7.3.

4. UMVU estimators

It may be natural to hope to find estimators that have a uniformly smallest risk in an interesting class of estimators. However, this short section describes a negative result which indicates that such uniformly best estimates do not exist in interesting cases. This fact justifies "weaker" optimality criteria such as the minimax, which considers the maximal risk rather than the risk at each value of the parameter, or the Bayes risk, which averages the risk over the parameter space. For further references and discussions see Cassel, Särndal, and Wretman (1977).

UMVU estimators are unbiased estimators whose MSE is smaller than that of any other unbiased estimator for each $\mathcal{Y} \in \Upsilon$. We consider more general risk functions than MSE but still use the term UMVU.

Definition 18 *A $\mathcal{P}$-unbiased estimator t^* (of a parameter θ) that is in some class of estimators, is said to be UMVU in this class, under the design $\mathcal{P}$, if*

$$R(\mathcal{P}, t^*; \mathcal{Y}) \leq R(\mathcal{P}, t; \mathcal{Y}) \quad \text{for all} \quad \mathcal{Y} \in \Upsilon \tag{25}$$

for any $\mathcal{P}$-unbiased estimator t of θ in this class.

We briefly discuss estimation of the population mean, and show that interesting cases of UMVU estimators do not exist; the condition that (25) hold for all $\mathcal{Y} \in \Upsilon$ is too strong.

Consider the so-called **generalized difference estimator** (Basu 1971) defined for any design $\mathcal{P}$ with inclusion probabilities $\alpha_i > 0$ for all $i \in \mathcal{N}$ by

$$t_{GD} = \sum_{i \in S} \frac{y_i - e_i}{\alpha_i} + \tilde{e}, \quad \text{where} \quad \tilde{e} = \sum_{i=1}^{N} e_i, \tag{26}$$

with known but arbitrary constants $\mathbf{e} = (e_1, \ldots, e_N)$. When $\mathbf{e} = \mathbf{0}$ we obtain the Horvitz-Thompson estimator. Note that for any $\mathbf{e}$ we have $\mathrm{E}_{\mathcal{P}}(t_{GD}/N) = \bar{\mathcal{Y}}$.

Proposition 19 *Let $\mathcal{P}$ be a design such that $\alpha_i > 0$ for all $i \in \mathcal{N}$ and $\alpha_i < 1$ for some $i \in \mathcal{N}$, and let $\theta = \bar{\mathcal{Y}}$. Let $\Upsilon = \Lambda^N$ with $\Lambda \subseteq \mathbb{R}$ such that $|\Lambda| \geq 2$. Consider a loss function $L(\tau, \theta)$ such that $L(\tau, \theta) \geq 0$ for all τ and θ, and $L(\tau, \theta) = 0$ if and only if $\tau = \theta$. Then no UMVU estimator in the class of unbiased estimators of the population mean θ exists.*

Proof If $\mathbf{e}$ happens to coincide with some $\mathcal{Y} \in \Upsilon$ then $t_{GD}/N = \bar{\mathcal{Y}}$ for any sample S, and $R(\mathcal{P}, t_{GD}/N; \mathcal{Y}) = 0$. It follows that a UMVU estimator t must satisfy $R(\mathcal{P}, t; \mathcal{Y}) = 0$ for all $\mathcal{Y} \in \Upsilon$. The result now follows readily. ■

Since t_{GD} is in the class of unbiassed linear (or affine) estimators (a linear combination of the observations plus a constant), the proof shows that there is no UMVU estimator in this class. This was shown by Godambe (1955). For further references see Godambe and Joshi (1965).

Finally we point out that *among symmetric unbiased estimates there do exist UMVU estimates* in a trivial manner. For example, the completeness result of Lemma 8 shows that $\bar{y}_S$ is the unique unbiased estimator of $\bar{\mathcal{Y}}$ that is symmetric, that is, an estimator of the form $t = t(\mathcal{Y}_S)$. Thus $\bar{y}_S$ is trivially UMVU among symmetric estimators. More generally, if we restrict attention to the class of symmetric unbiased estimators of any parameter, then at most one such estimator exists, and it is trivially UMVU in this class.

5. Admissibility

Definition 20 *A strategy* $(\mathcal{P}_0, t_0)$ *is* **admissible** *in a class of strategies if there is no strategy* $(\mathcal{P}, t)$ *in this class satisfying*

$$R(\mathcal{P}, t; \mathcal{Y}) \leq R(\mathcal{P}_0, t_0; \mathcal{Y}) \text{ for all } \mathcal{Y} \in \Upsilon \text{ with strict inequality for at least one } \mathcal{Y}.$$

An estimator t_0 *is* **admissible** *under a design* $\mathcal{P}_0$ *in a class of estimators if there is no estimator* t *in this class satisfying*

$$R(\mathcal{P}_0, t; \mathcal{Y}) \leq R(\mathcal{P}_0, t_0; \mathcal{Y}) \text{ for all } \mathcal{Y} \in \Upsilon \text{ with strict inequality for at least one } \mathcal{Y}.$$

If the first inequality in the above definition holds, we say that the strategy $(\mathcal{P}, t)$ **dominates** $(\mathcal{P}_0, t_0)$, and if the second inequality holds, we say that t **dominates** t_0 under $\mathcal{P}_0$.

Admissibility is in some sense a minimal property. If a strategy (estimator) is inadmissible, then there is a better strategy (estimator) that will perform better (or at least as well) under any of the criteria mentioned in this chapter. But an admissible strategy may still be very poor. For example, it is easy to construct a finite population estimation problem such that an estimator which is a constant guess that ignores the sample altogether is admissible in a wide class, but has an arbitrarily large risk on large parts of the parameter space. Admissibility is called *Pareto optimality* in the terminology of game theory.

The next two theorems, from Scott (1975), shows that admissibility is a property of the support of the design $\mathcal{P}$ defined by $\mathcal{S}_{\mathcal{P}} = \{S \subseteq \mathcal{N} \; : \; \mathcal{P}(S) > 0\}$. In fact, if t_0 is admissible under a design $\mathcal{P}$, then it is admissible under any design having the same or smaller support. Here convexity of the loss function and randomized rules play a role, as will be seen in the precise statements and proofs.

Theorem 21 *If an estimator* t_0 *is admissible under a design* $\mathcal{P}_0$ *in the class of all estimators including randomized ones, then the same holds for any design* $\mathcal{P}$ *satisfying* $\mathcal{S}_{\mathcal{P}} \subseteq \mathcal{S}_{\mathcal{P}_0}$.

Proof Suppose to the contrary that there exists an estimator t_1 that dominates t_0 under $\mathcal{P}$. Using it, we will construct an estimator t^* that dominates t_0 under $\mathcal{P}_0$, contradicting our assumption.

Let $m = \max\{\mathcal{P}(S)/\mathcal{P}_0(S) \; : \; S \in \mathcal{S}_{\mathcal{P}}\}$ and $Q(S) = \mathcal{P}(S)/m\mathcal{P}_0(S)$. Then by the conditions on the supports $1 \leq m < \infty$ and $0 < Q(S) \leq 1$. Here and in the next proof we use the abbreviation $t(S)$ for $t(S, \mathcal{Y})$, that is, we suppress the parameter $\mathcal{Y}$. Consider

the following randomized estimator t^*: if $S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}$ then $t^*(S) = t_0(S)$. If $S \in \mathcal{S}_{\mathcal{P}}$ then $t^*(S) = t_1(S)$ w.p. $Q(S)$ and $t^*(S) = t_0(S)$ w.p. $1 - Q(S)$. We claim that t^* dominates t_0 under $\mathcal{P}_0$. Indeed,

$$
\begin{aligned}
&R(\mathcal{P}_0, t^*; \mathcal{Y}) \\
={} & \sum_{S \in \mathcal{S}_{\mathcal{P}}} Q(S)\mathcal{P}_0(S)L(t_1(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
={} & m^{-1} \sum_{S \in \mathcal{S}_{\mathcal{P}}} \mathcal{P}(S)L(t_1(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
\leq{} & m^{-1} \sum_{S \in \mathcal{S}_{\mathcal{P}}} \mathcal{P}(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
={} & \sum_{S \in \mathcal{S}_{\mathcal{P}}} Q(S)\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
={} & R(\mathcal{P}_0, t_0; \mathcal{Y})
\end{aligned}
$$

for all $\mathcal{Y}$, with a strict inequality for at least one $\mathcal{Y}$, where the inequality follows from the assumption that t_1 dominates t_0 under $\mathcal{P}$. Note that the estimator t^* is randomized. If $L(\tau, \mathcal{Y})$ is convex in τ we can replace t^* by its expectation, and thus assume that t^* is nonrandomized. See Remark 4. In this case Theorem 21 holds also for the class of nonrandomized estimators. ∎

For the next result we slightly generalize Scott's (1975) formulation. Given a collection H of real valued functions defined on subsets of $\mathcal{N}$ and a corresponding collection of constants $C = \{c_h\}_{h \in H}$, define the class of designs

$$\mathcal{D}_{H,C} = \{\mathcal{P} : E_{\mathcal{P}} h(S) = c_h \text{ for all } h \in H\}.$$

For H consisting of the single function $h(S) = |S|$, and $c_h = n$ we obtain the class of designs with expected sample size $= n$. Taking H to be the set of indicator functions of all sets of size $\neq n$, and $c_h = 0$, we obtain the class of design having fixed sample size n. These two classes were considered by Scott.

Given an estimator $t_0(S, \mathcal{Y})$ and parameter $\theta = \theta(\mathcal{Y})$, the class of designs $\mathcal{P}$ under which t_0 is unbiased for θ is also of this kind. To see this define $h_{\mathcal{Y}}(S) = t_0(S, \mathcal{Y})$, and $c_h = c_{\mathcal{Y}} = \theta(\mathcal{Y})$, and set $H = \{h_{\mathcal{Y}} : \mathcal{Y} \in \Upsilon\}$.

The class of designs of having sample size n in some interval, the class of conditional SRS designs, see definition following (9), and other classes are of the above type, as are their intersections.

Theorem 22 *If a strategy $(\mathcal{P}_0, t_0)$ is admissible in a class of strategies having designs in a given class $\mathcal{D} = \mathcal{D}_{H,C}$ and estimators in the class of all estimators including randomized ones, then the same holds for any strategy $(\mathcal{P}, t_0)$ such that $\mathcal{P} \in \mathcal{D}$, and $\mathcal{S}_{\mathcal{P}} \subseteq \mathcal{S}_{\mathcal{P}_0}$.*

Proof For $\mathcal{P}$ as above, suppose to the contrary that there exist a design $\mathcal{P}_1$ in $\mathcal{D}$, and an estimator t_1, such that the strategy $(\mathcal{P}_1, t_1)$ dominates $(\mathcal{P}, t_0)$. Set $\mathcal{P}^*(S) = \mathcal{P}_0(S) + m^{-1}(\mathcal{P}_1(S) - \mathcal{P}(S))$, and note that $\mathcal{P}^* \in \mathcal{D}$. Define $T(S) = \mathcal{P}_1(S)/m\mathcal{P}^*(S)$. Since $\mathcal{P}_0(S) - m^{-1}\mathcal{P}(S) \geq 0$, we have $\mathcal{P}^*(S) \geq m^{-1}\mathcal{P}_1(S)$, and therefore $0 \leq T(S) \leq 1$. Define the randomized estimator $t^*(S) = t_1(S)$ with probability $T(S)$ and $t^*(S) = t_0(S)$ with probability $1 - T(S)$. A calculation similar to the one in the proof of Theorem 21 shows that $R(\mathcal{P}^*, t^*; \mathcal{Y})$ dominates $R(\mathcal{P}_0, t_0; \mathcal{Y})$, a contradiction. ∎

Godambe and Joshi (1965) have shown that for any design, the Horvitz-Thompson estimator is admissible in the class of all unbiased estimators of a finite population total. The proof we give is essentially due to Ramakrishnan (1973), extended here from MSE to a more general convex loss function. The requirement $0 \in \Lambda$ is discussed after the proof.

Theorem 23 *Let $\mathcal{P}$ be any design with $\alpha_i > 0$ for $i = 1, 2, \ldots, N$, and consider the parameter space Λ^N for some set Λ satisfying $0 \in \Lambda$. The Horvitz-Thompson estimator $t_{HT}(S, \mathcal{Y}) = \sum_{i \in S} y_i/\alpha_i$ is admissible in the class of unbiased estimators for the parameter $\theta_N = \sum_{i=1}^N y_i$ provided that the loss function $L(t, \theta)$ is strictly convex in t and assumes its minimum when $t = \theta$.*

Proof We assume $\mathcal{P}(S) > 0$ implies $|S| > 0$ to avoid trivialities. The proof is by induction on N. For $N = 1$ clearly $t_{HT} = \theta_1$, and the result is obvious. The induction hypothesis is that for a population of size N, $R(\mathcal{P}, t; \mathcal{Y}) \leq R(\mathcal{P}, t_{HT}; \mathcal{Y})$ for all $\mathcal{Y} \in \Lambda^N$ implies that $t = t_{HT}$ with $\mathcal{P}$-probability 1, and it is easy to see that the desired admissibility follows. Let $(\mathcal{P}^*, t^*)$ be an unbiased strategy for θ_{N+1} on a population of size $N + 1$ denoted by U_{N+1}, and consider the population U_N of size N obtained by removing the last coordinate from U_{N+1}. On the latter population we construct a strategy $(\mathcal{P}, t)$ by setting

$$\mathcal{P}(S) = \mathcal{P}^*(S) + \mathcal{P}^*(S, N+1), \quad t(S, \mathcal{Y}) = \frac{1}{\mathcal{P}(S)}[\mathcal{P}^*(S)t^*(S, \mathcal{Y}^*) + \mathcal{P}^*(S, N+1)t^*((S, N+1), \mathcal{Y}^*)]$$

where $(S, N+1) = S \cup \{N+1\}$, $\mathcal{Y} = (y_1, \ldots, y_N)$, and $\mathcal{Y}^* = (y_1, \ldots, y_N, 0)$. It is easy to see that $(\mathcal{P}, t)$ is unbiased for θ_N. For now let t_{HT} and t_{HT}^* denote the Horvitz-Thompson estimators for the designs $\mathcal{P}$ and $\mathcal{P}^*$ respectively. We claim that

$$R(\mathcal{P}, t_{HT}; \mathcal{Y}) = R(\mathcal{P}^*, t_{HT}^*; \mathcal{Y}^*). \tag{27}$$

To see this, construct t_{HT} and t_{HT}^* on the same probability space (coupling) as follows. When a set $S \subset \{1, \ldots, N\}$ or $(S, N+1)$ is chosen with probability $\mathcal{P}^*$ as the sample for t_{HT}^*, let S be the chosen set for t_{HT}. It then follows that $t_{HT} = t_{HT}^*$, and (27) follows.

Next we claim that

$$R(\mathcal{P}, t; \mathcal{Y}) \le R(\mathcal{P}^*, t^*; \mathcal{Y}^*). \tag{28}$$

It is easy to see that this follows by the convexity of L and Jensen's inequality, given that t is a convex combination of values of t^*. Moreover, the fact that L is strictly convex implies strict inequality in (28) whenever $t^*(S, \mathcal{Y}^*) \ne t^*((S, N+1), \mathcal{Y}^*)$, $\mathcal{P}^*(S) > 0$, and $\mathcal{P}^*(S, N+1) > 0$.

Assume that $R(\mathcal{P}^*, t^*; \mathcal{Y}^*) \le R(\mathcal{P}^*, t^*_{HT}; \mathcal{Y}^*)$ for all $\mathcal{Y}^* \in \Lambda^{N+1}$. Together with (27) and (28) we then have

$$R(\mathcal{P}, t; \mathcal{Y}) \le R(\mathcal{P}, t_{HT}; \mathcal{Y}) \text{ for all } \mathcal{Y} \in \Lambda^N \tag{29}$$

and so by the induction hypothesis

$$t(S, \mathcal{Y}) = t_{HT}(S, \mathcal{Y}) \tag{30}$$

for any S with $\mathcal{P}(S) > 0$. For such sets $S \subseteq \{1, \ldots, N\}$ we clearly have

$$t_{HT}(S, \mathcal{Y}^*) = t^*_{HT}(S, \mathcal{Y}^*) \text{ for all } \mathcal{Y}^* \in \Lambda^{N+1}. \tag{31}$$

Moreover, strict inequality would hold in (29) for any $\mathcal{Y}$ such that $t^*(S, \mathcal{Y}^*) \ne t^*((S, N+1), \mathcal{Y}^*)$, $\mathcal{P}^*(S) > 0$, and $\mathcal{P}^*(S, N+1) > 0$ for $\mathcal{Y}^* = (y_1, \ldots, y_N, 0)$. But strict inequality is impossible since it would contradict the induction hypothesis, and therefore if $\mathcal{P}^*(S) > 0$ then either $\mathcal{P}^*(S, N+1) = 0$ or $t^*(S, \mathcal{Y}^*) = t^*((S, N+1), \mathcal{Y}^*)$. In either case we then have $t(S, \mathcal{Y}) = t^*(S, \mathcal{Y}^*)$. This, together with (30) and (31), implies that

$$t^*(S^*, \mathcal{Y}^*) = t^*_{HT}(S^*, \mathcal{Y}^*) \text{ for all } \mathcal{Y}^* \in \Lambda^{N+1} \tag{32}$$

for any S^* not containing $N+1$ such that $\mathcal{P}^*(S) > 0$. We can repeat the argument with the label $N+1$ replaced by any j, and obtain (32) for any set S^* of size $\le N+1$. Finally, (32) for the set $S^* = \{1, \ldots, N+1\}$ follows from this equality for all other sets S^* and from the fact that t^* and t^*_{HT} have the same expectation. This completes the induction step. ∎

The above result required $0 \in \Lambda$. Unlike in the case of Proposition 17, we cannot assume it "without loss of generality" by applying a linear transformation when $0 \notin \Lambda$. Indeed, if $\Lambda = \{a\}$ with $a \ne 0$ then the estimator $t = a$ is better than t_{TH} with respect to any design such that $\mathrm{Var}\sum_{i \in S} 1/\alpha_i > 0$. It is easy to construct less trivial examples. However, for

$A = \{0, 1\}$ and using t_{HT}/N for estimating a proportion, the above admissibility result holds.

It is easy to construct examples with fixed or random sample size, where the Horvitz-Thompson estimator t_{HT}/N for a proportion is not in the interval $[0, 1]$ with positive probability (a trivial example is $n = 1$, $0 < \alpha_i < 1$ and $y_i \equiv 1$). In this case, it is clearly not admissible in the class of all estimators. This does not contradict Theorem 23 which requires unbiasedness (and allows designs with random sample size). For fixed-size sample designs, the Horvitz-Thompson estimator is admissible among all estimators when the parameter space is $\mathbb{R}^N$; see Joshi (1965, 1966). It follows that for SRS of any size n, the sample mean is an admissible estimator of the population mean. By Theorem 21 it follows that the sample mean is admissible for any fixed-size design.

The following example shows that if the sample size is random, t_{HT} may not be admissible in the class of all estimators: set $N = 2$ and $\mathcal{P}(\{1\}) = \mathcal{P}(\{1, 2\}) = 1/2$. When the sample $\{1, 2\}$ is selected, we have $t_{HT} = y_1 + 2y_2$, and the (biased) estimator obtained by instead using $y_1 + y_2$ shows that t_{HT} is not admissible.

6. Superpopulation models

6.1. Background

In ***superpopulation models*** one assumes that the given population $\mathcal{Y} = (y_1, \ldots, y_N)$ is a realization of a random vector $Y = (Y_1, \ldots, Y_N)$ having a distribution $\mathcal{G}$. We shall refer to $\mathcal{G}$ as the ***prior***. Several possibilities arise: **1.** $\mathcal{G}$ is completely known. **2.** $\mathcal{G}$ belongs to a class having some known parameters and properties, e.g., distributions with certain specified moments and possibly with some exchangeability properties. **3.** $\mathcal{G}$ depends on a parameter ϕ, that is, $\mathcal{G} = \mathcal{G}_\phi$.

Design-based inference on the population $\mathcal{Y}$, as the name suggests, uses the sampling design (randomization distribution) only. Pure *model-based inference* on the population $\mathcal{Y}$, the prior $\mathcal{G}$, or the parameter ϕ, refers to inference where the sampling design plays no role, and the risk, for example is defined as expectation with respect to $\mathcal{G}$ of the squared difference between the estimate and the estimand, conditioned on the sample.

A third approach combines the above two. Starting from the design-based risk $R(\mathcal{P}, t; Y)$, this approach studies the Bayes risk (see definition 3), that is, the expected risk with respect to $\mathcal{G}$, $\; E_{\mathcal{G}} R(\mathcal{P}, t; Y)$. The optimization goal is to find a strategy $(\mathcal{P}, t)$ that minimizes the latter expectation. For unbiased estimators and quadratic loss, this expectation becomes $E_{\mathcal{G}} Var_{\mathcal{P}} t$, known in the sampling literature as the *anticipated variance*. It is often used to compare two design unbiased estimators when comparison of the $\mathcal{P}$-variances does not lead to clear conclusions.

It may happen that the superpopulation assumptions involve enough symmetry and randomness to make the sampling design inessential. For example, if Y is exchangeable under the superpopulation model and we use a symmetric estimator, then random sampling may be redundant since the data are assumed to be given in a random order.

We have already used the Bayesian approach, and, in fact, (22) can be seen as a prior of the above type; however, we used it only as a technical device to arrive at a minimax estimator, noting that the minimax criterion does not depend on the Bayesian structure.

We shall not discuss the philosophy and relevance of superpopulation models and model-based optimality criteria here. Some discussions and references can be found, for example, in Smith (1976), Särndal, Swensson, and Wretman(1992), Hedayat and Sinha (1991), and Cassel, Särndal, and Wretman (1977); the latter two books also contain a discussion that is closely related to the one that follows, with references and further results.

6.2. $\mathcal{P}$-unbiased estimators

In the discussion below we consider $\mathcal{P}$-***unbiased estimators*** of the ***population mean*** $\bar{\mathcal{Y}}$. We shall consider ***quadratic loss*** and MSE, and the Bayes risk, which is the MSE integrated with respect to the prior $\mathcal{G}$. Theorem 30 shows that for any exchangeable prior (superpopulation model) the Bayes risk is minimized among $\mathcal{P}$-unbiased strategies by the strategy consisting of SRS (or any design with $\alpha_i = n/N$) and the sample mean. This is generalized in Theorem 31 to the case of exchangeability of a linear transformation of the population values, and the optimal estimators are then the generalized difference estimators (see (26)) of which Horvitz-Thompson estimators form a special case. Note that these results involve $\mathcal{P}$-unbiasedness, which is a design-based criterion, and the Bayes risk, which is a model-based expectation over a design-based risk.

The results and techniques used next: sufficiency, completeness and the Rao-Blackwell approach are close to those that led to Theorem 9. However, many detail are different. In particular, here the notions of sufficiency and completeness are with respect to the prior $\mathcal{G}$ rather than the design as in Section 3.2.

When we think of the population as fixed we denote it by $\mathcal{Y}$; when we want to emphasize that under the superpopulation model it is random we denote it by Y. We used $\bar{\mathcal{Y}}$ and $\bar{y}_S$ for the population and sample means; we denote them by $\bar{Y}$ and $\bar{Y}_S$, when we want to emphasize that now the population is random, and when we take expectation with respect to $\mathcal{G}$. Given $Y = (Y_1, \ldots, Y_N)$ and a sample S, Y_S denotes the ***multiset*** containing all Y_i-values arising from distinct labels $i \in S$, in analogy to $\mathcal{Y}_S$ in the fixed population case. Similarly, we may express the data $D[S, \mathcal{Y}_S])$ as $D[S, Y_S])$, and when we want to describe an estimator, we may write $t(S, Y_S)$ instead of $t(S, \mathcal{Y}_S)$, etc.

Note that we now have two sources of randomness, the sample $S \sim \mathcal{P}$ and the population

$Y \sim \mathcal{G}$. Therefore, notations like $E_{\mathcal{P}}$, $E_{\mathcal{G}}$, and $E_{\mathcal{G},\mathcal{P}}$ for expectations will be used, where $E_{\mathcal{G},\mathcal{P}} = E_{\mathcal{G}}E_{\mathcal{P}}$. Unless otherwise stated, we consider designs that are **noninformative** or **ignorable**, that is $\mathcal{P}(S|Y) = \mathcal{P}(S)$, independent of Y. In words, the design does not depend on the (initially known) population values Y. This assumption allows interchange of expectations with respect to $\mathcal{G}$ and $\mathcal{P}$. We discuss it further in Section 6.3.

Recall that the strategy $(\mathcal{P}, t)$ is **unbiased** for $\bar{\mathcal{Y}}$ (the population mean) if t is $\mathcal{P}$-**unbiased**, that is, if for all $\mathcal{Y} \in \Upsilon$, $E_{\mathcal{P}}t := \sum_S \mathcal{P}(S)t(D[S, \mathcal{Y}_S]) = \bar{\mathcal{Y}}$. Note that the latter expectation can also be interpreted as the conditional expectation $E\{t(D[S, Y_S]) \,|\, Y = \mathcal{Y}\}$. Recall also that for S satisfying $|S| = n$, $\bar{Y}_S = \frac{1}{n}\sum_{i \in S} Y_i$, that is, $\bar{Y}_S = \frac{1}{|S|}\sum_{i \in S} Y_i$.

Let $\mathbb{G}$ denote the class of **exchangeable distributions**, that is, distributions that remain unchanged under permutations of the components of the vector Y.

Lemma 24 *Let $(\mathcal{P}, t = t(D[S, \mathcal{Y}_S]))$ be an unbiased strategy for $\bar{\mathcal{Y}}$. Let $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G} \in \mathbb{G}$, and $E_{\mathcal{G}}Y_i = \mu_{\mathcal{G}}$. Then $E_{\mathcal{G},\mathcal{P}}t(D[S, Y_S]) := E_{\mathcal{G}}\sum_S \mathcal{P}(S)t(S, Y_S) = \mu_{\mathcal{G}}$. Also, $E_{\mathcal{G},\mathcal{P}}\bar{Y}_S = \mu_{\mathcal{G}}$.*

Proof The first part of the lemma is obvious. For the second part, note that for a general sampling design $\mathcal{P}$, $\bar{Y}_S$ is not necessarily $\mathcal{P}$-unbiased, so the first part does not imply the second. We have $\bar{Y}_S = \sum_{i=1}^{N} Y_i I_i / \sum_{j=1}^{N} I_j$ where $I_i = 1$ if $i \in S$ and 0 otherwise. Now $E_{\mathcal{G}}(\bar{Y}|S) = \mu_{\mathcal{G}}\sum_{i=1}^{N} I_i / \sum_{j=1}^{N} I_j = \mu_{\mathcal{G}}$, and the result follows. ■

The next two easy lemmas show completeness and sufficiency. The classical Rao-Blackwell argument uses completeness and sufficiency as follows: given a statistic $t(X)$ which depends on some data $X \sim P_{\theta}$ (see Definition 3), and a sufficient statistic for θ, say $W(X)$, the estimator $t_0 = E(t|W)$ is a statistic since it does not depend on θ by sufficiency. Also, t_0 has the same expectation as t, but a smaller variance (by Jensen's inequality, or by a well-known variance decomposition formula). If W is complete, then t_0 is the unique estimator with the same expectation as t. This proves that it is a UMVU estimator of Et (see Definition 3). A version of this argument appears below, leading to Theorem 30.

Lemma 25 *Let $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G} \in \mathbb{G}$ and let $S \sim \mathcal{P}$. Consider the data $D = D[S, Y_S]$. Then Y_S is sufficient in the sense that $P(D|Y_S)$ does not depend on $\mathcal{G}$. (It does depend on $\mathcal{P}$, which is held fixed here.)*

Proof Just note that if $|S| = n$, then $P(D|Y_S) = \mathcal{P}(S)/n!$, where the $n!$ is due to the $n!$ equally likely (by exchangeability) ways of pairing the elements of S with those of Y_S. ■

For the next lemma we need two new conditions, which will henceforth be assumed. The first is that the **parameter space is a product** of the form $\Upsilon = \Lambda^N$, and the second is that the design $\mathcal{P}$ has a **fixed sample size**, say n.

Lemma 26 *Let $\mathbb{G}$ denote the class of exchangeable distributions over a product space Λ^N, and let $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G} \in \mathbb{G}$, and $S \sim \mathcal{P}$, a given design with fixed sample size n. Let Y_S denotes the multiset containing all Y_i-values arising from distinct labels $i \in S$. Then Y_S is complete; that is, for any symmetric (permutation invariant) function h of n variables, if $E_{\mathcal{G},\mathcal{P}} h(Y_S) = 0$ for all $\mathcal{G} \in \mathbb{G}$ then $P_{\mathcal{G},\mathcal{P}}(h(Y_S) \neq 0) = 0$ for all $\mathcal{G} \in \mathbb{G}$.*

Proof The proof is similar to that of Lemma 8. For any $a \in \Lambda$ let $\mathcal{G}$ be the probability measure concentrated on $(a, \ldots, a) \in \Lambda^N$. Then clearly for this $\mathcal{G}$, $E_{\mathcal{G},\mathcal{P}} h(Y_S) = 0$ implies $h(a, \ldots, a) = 0$. Now let $\mathcal{G}$ be the exchangeable probability measure which concentrates on $(b, a, \ldots, a) \in \Lambda^N$ and all its permutations. This is used to prove $h(b, a, \ldots, a) = 0$ as in the proof of Lemma 8, and so on. ∎

As usual, completeness implies uniqueness of unbiased estimators, since if there existed two distinct unbiased estimators which are functions of Y_S, then their difference h would be a non zero function whose expectation is zero, contradicting Lemma 26. The following example shows that a fixed sample size is indeed needed in Lemma 26. If the sample size is random with expectation n, then it is easy to see that the estimator $t = \frac{1}{n} \sum_{i \in S} Y_i$, where we divide by the expected sample size n rather than $|S|$, satisfies $E_{\mathcal{G},\mathcal{P}} t = \mu_{\mathcal{G}}$. The same holds for the estimator $\bar{Y}_S = \frac{1}{|S|} \sum_{i \in S} Y_i$ by Lemma 24, and unless the sample size is fixed, we have two distinct unbiased estimators of $\mu_{\mathcal{G}}$.

We shall now consider **quadratic loss**. Then $R(\mathcal{P}, t; Y) = \sum_S \mathcal{P}(S)(t(S, Y_S) - \bar{Y})^2 = Var_{\mathcal{P}} t$, and $E_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2 = E_{\mathcal{G}} \sum_S \mathcal{P}(S)(t(S, Y_S) - \mu_{\mathcal{G}})^2$, where the sum extends over all subsets S (of size n) of $\mathcal{N}$. The following lemma shows that for unbiased estimation of $\mu_{\mathcal{G}}$, the sample mean $\bar{Y}_S$ is optimal in the sense of minimizing $E_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2$. In fact, Lemma 27 holds for any convex loss function, but since quadratic loss is required in all the subsequent lemmas and theorems, we use quadratic loss also here.

Lemma 27 *Let $\bar{Y}_S = \frac{1}{n} \sum_{i \in S} Y_i$, and let $t = t(D[S, Y_S])$ be an estimator satisfying $E_{\mathcal{G},\mathcal{P}} t(D[S, Y_S]) = \mu_{\mathcal{G}}$. Then*

$$E_{\mathcal{G},\mathcal{P}}(\bar{Y}_S - \mu_{\mathcal{G}})^2 \leq E_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2. \tag{33}$$

In particular this holds for any $\mathcal{P}$-unbiased estimator t of the population mean $\bar{\mathcal{Y}}$.

Proof The lemma follows by a standard Rao-Blackwell argument applied to the quadratic loss function, and using the facts that Y_S is sufficient and complete for the parameter $\mathcal{G}$, and that $E_{\mathcal{G},\mathcal{P}} \bar{Y}_S = \mu_{\mathcal{G}}$. The latter equality and the last part about $\mathcal{P}$-unbiased estimators follow from Lemma 24. ∎

The next lemma is a standard variance decomposition. Recall the notation $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$, and the fact that for a $\mathcal{P}$-unbiased estimator of $\bar{\mathcal{Y}}$ we have $E_{\mathcal{P}}(t|Y) = \bar{Y}$.

Lemma 28 *Let t be a $\mathcal{P}$-unbiased estimator of $\bar{\mathcal{Y}}$. Then $Var_{\mathcal{G},\mathcal{P}}\, t := E_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2 = E_{\mathcal{G}}Var_{\mathcal{P}}t + E_{\mathcal{G}}(\bar{Y} - \mu_{\mathcal{G}})^2 = E_{\mathcal{G}}\sum_S \mathcal{P}(S)(t(S, Y_S) - \bar{Y})^2 + E_{\mathcal{G}}(\bar{Y} - \mu_{\mathcal{G}})^2.$*

Lemmas 27 and 28 imply

Lemma 29 *Let $\mathcal{P}$ be any design with fixed sample size n, $t = t(D[S, \mathcal{Y}_S])$ a $\mathcal{P}$-unbiased estimator of $\bar{\mathcal{Y}}$, and let $\mathcal{G}$ be any exchangeable (prior) distribution on the population $Y = (Y_1, \ldots, Y_N)$. Then*

$$E_{\mathcal{G}}Var_{\mathcal{P}}\bar{Y}_S = E_{\mathcal{G}}R(\mathcal{P}, \bar{Y}_S; Y) \le E_{\mathcal{G}}Var_{\mathcal{P}}t = E_{\mathcal{G}}R(\mathcal{P}, t; Y).$$

The above result compares a $\mathcal{P}$-unbiased estimator t to the estimator $\bar{Y}_S$, which in general is not $\mathcal{P}$-unbiased. In fact, the strategy $(\mathcal{P}, \bar{Y}_S)$ is unbiased if and only if $\alpha_i = n/N$. Note that $E_{\mathcal{G}}Var_{\mathcal{P}}\bar{Y}_S = E_{\mathcal{G}}\sum_S \mathcal{P}(S)(\bar{Y}_S - \bar{Y})^2 = E_{\mathcal{G}}\sum_\pi \mathcal{P}(\pi S)(\bar{Y}_{\pi S} - \bar{Y})^2$ is constant as a function of $\mathcal{P}$ for all designs having sample size n, since by exchangeability $\bar{Y}_{\pi S}$ are identically distributed for all permutations π. Thus we obtain the following theorem that compares the Bayes risk of unbiased strategies:

Theorem 30 *Any strategy $(\mathcal{P}_0, \bar{Y}_S)$ with fixed sample size n, and $\alpha_i = n/N$, is optimal in the class of $\mathcal{P}$-unbiased (for the population mean) strategies $(\mathcal{P}, t = t(D[S, Y_S]))$ having sample size n, in the sense that for any $\mathcal{G} \in \mathbb{G}$,*

$$E_{\mathcal{G}}R(\mathcal{P}_0, \bar{Y}_S; Y) \le E_{\mathcal{G}}R(\mathcal{P}, t; Y). \tag{34}$$

The above result can be generalized as follows. Suppose that we have reason to believe that our units are not exchangeable. For example, they may have different known average sizes a_i, that is, $\mu_i := E_{\mathcal{G}}Y_i = a_i$ and, more generally, $\mu_i = E_{\mathcal{G}}Y_i = a_i\mu + b_i$ and perhaps also $E_{\mathcal{G}}(Y_i - \mu_i)^2 = a_i^2\sigma^2$. The known constants a_i, b_i can be viewed as auxiliary information. This leads to Theorem 31 below, in which we assume that the variables $(Y_1 - b_1)/a_1, \ldots, (Y_N - b_N)/a_N$ have an exchangeable prior (superpopulation model) with known constants $a_i > 0$ and b_i. We set $\sum_{i=1}^N a_i = N$ without loss of generality.

Theorem 31 *Let $((Y_1 - b_1)/a_1, \ldots, (Y_N - b_N)/a_N) \sim \mathcal{G} \in \mathbb{G}$, and*

$$t_{GD_0} = \sum_{i \in S} \frac{Y_i - b_i}{\alpha_i} + \tilde{b}, \quad where \ \tilde{b} = \sum_{i=1}^N b_i.$$

Let $\mathcal{P}_0$ be any design having a fixed sample size n, and $\alpha_i = a_i n/N$. The strategy $(\mathcal{P}_0, \frac{1}{N}t_{GD_0})$ is optimal in the class of $\mathcal{P}$-unbiased (for the population mean) strategies

$(\mathcal{P}, t = t(D[S, Y_S]))$ *having sample size n, in the sense that*

$$E_{\mathcal{G}} R(\mathcal{P}_0, \tfrac{1}{N} t_{GD_0}; Y) \le E_{\mathcal{G}} R(\mathcal{P}, t; Y). \tag{35}$$

Proof Define $Z_i = \frac{(Y_i - b_i)}{a_i} + \bar{b}$, where $\bar{b} = \frac{1}{N} \sum_{i=1}^{N} b_i$, and $Z = (Z_1, \dots, Z_N)$. Then $\bar{Z}_S := \frac{1}{n} \sum_{i \in S} Z_i = \frac{1}{N} t_{GD_0}$. The proof is the same as that of Theorem 30, applied to the above Z. ∎

Theorem 31 is due to Cassel, Särndal, and Wretman (1977). The special case of $b_i = 0$ shows that Horvitz-Thompson strategies, that is, any strategy $(\mathcal{P}_0, \frac{1}{N} t_{HT})$ with $\alpha_i = a_i n / N$ and the corresponding estimate $\frac{1}{N} t_{HT} = \frac{1}{N} \sum_{i \in S} y_i / \alpha_i$, have a minimal Bayes risk among $\mathcal{P}$-unbiased (for the population mean) strategies for priors such that the vector $(Y_1/a_1, \dots, Y_N/a_N)$ is exchangeable. In this case, the expectations EY_i are proportional to some known constants, and any design of fixed sample size n with inclusion probabilities that are proportional to those constants and a corresponding Horvitz-Thompson estimator form an optimal strategy with respect to Bayes risk.

6.3. Linear prediction

We consider estimation of the population mean on the basis of a sample from the random population Y, where $Y \sim \mathcal{G}$, to be specified later. Under such a superpopulation model, the population mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ is a random variable which we are trying to predict. From the relation $\bar{Y} = \frac{n}{N} \bar{Y}_S + (1 - \frac{n}{N}) \bar{Y}_{S^c}$ we see that when $D[S, Y_S]$ is observed, our task is to predict $\bar{Y}_{S^c} = \frac{1}{N-n} \sum_{i \in S^c} Y_i$, where $n = |S|$ and S^c denotes the complement of S. We consider momentarily the possibility that the design depends on the population values, in which case the ***design*** is said to be ***informative***, and write $\mathcal{P}(S|Y)$ for the probability of sampling S given that the population vector is Y. Let $g(y)$, $y \in \mathbb{R}^N$ denote a density of the prior $\mathcal{G}$ (which may depend on a parameter). Then the ***predictive density*** of Y_{S^c}, the unobserved part of Y given the data, is

$$f(y_{S^c}|S, y_S) = \mathcal{P}(S|y_S, y_{S^c}) g(y_S, y_{S^c}) \Big/ \int \mathcal{P}(S|y_S, y_{S^c}) g(y_S, y_{S^c}) dy_{S^c}.$$

The design is ***noninformative*** or **ignorable** if $\mathcal{P}(S|Y) = \mathcal{P}(S)$, independent of Y. ***Adaptive designs*** satisfy $\mathcal{P}(S|Y) = \mathcal{P}(S|Y_S)$; that is, the sample may depend on the observed y-values, but not on the unobserved ones. Such designs arise when the sample is selected sequentially: first some units are sampled, and the choice of the units that are added to the sample depends on the y's already observed; see Thompson and Seber

(1996) and references therein. In either the ignorable or the adaptive case we obtain

$$f(y_{S^c}|S, y_S) = g(y_S, y_{S^c}) / \int g(y_S, y_{S^c}) dy_{S^c},$$

and we see that the design does not play a role in the predictive density. It is therefore not surprising that the predictive (model-based, Bayes) optimality result of Theorem 32 below is not stated in terms of a sampling design.

We need some definitions. A statistic $t = t(S, Y_S)$ is said to be a **linear predictor** (or estimator) if it is of the form $t(S, Y_S) = \sum_{i \in S} r_i Y_i + q$, where the constants r_i, q may depend on S.

Let s be a given subset of $\mathcal{N}$. Given a statistic $t = t(S, Y_S)$ we can fix the set s, and consider the random variable $t(s, Y_s)$ for the given fixed set s and $Y \sim \mathcal{G}$. The estimator t is said to be a $\mathcal{G}$-**unbiased predictor** of the population mean $\bar{Y}$ if $E_{\mathcal{G}}(t(s, Y_s) - \bar{Y}) = 0$ for **every** fixed $s \subset \mathcal{N}$. See, e.g., Cassel, Särndal, and Wretman (1977). $\mathcal{G}$-unbiased predictors of other parameters are defined similarly.

The above definition and the theorem below are written in terms of a fixed set (or nonrandom sample) s and expectations in the form $E_{\mathcal{G}}[\,\cdot\,]$, taken with respect to $Y \sim \mathcal{G}$. An equivalent formulation would be to consider a random S and replace these expectations by $E_{\mathcal{G}}[\,\cdot\,|S = s]$ provided that Y and S are independent, which means that the design $\mathcal{P}$ is ignorable. If $E_{\mathcal{G}}[(t(s, Y_s) - \bar{Y})|S = s] = 0$ for an ignorable design $\mathcal{P}$, then we can now take expectation with respect to $S \sim \mathcal{P}$, to obtain $E_{\mathcal{P}, \mathcal{G},}(t - \bar{Y}) = 0$. Exchanging the order of the expectations, it is easy to see that a $\mathcal{G}$-unbiased predictor t satisfies $E_{\mathcal{G}, \mathcal{P}}(t - \bar{Y}) = 0$ for any *ignorable* design $\mathcal{P}$. These operations cannot be done if $\mathcal{P}$ is informative, that is, if it depends on Y.

On the other hand, for any design $\mathcal{P}$ (ignorable or not), a $\mathcal{P}$-unbiased estimate t of $\bar{Y}$ satisfies $E_{\mathcal{G}, \mathcal{P}}(t - \bar{Y}) = 0$ for any $\mathcal{G}$.

Theorem 32 below, which is one of many results on optimality in the class of $\mathcal{G}$- unbiased predictors, appears in Hedayat and Sinha (1991) with further references. A closely related result appears in Royall (1970b). The auxiliary variables x_i, b_i below are assumed to be known constants.

Theorem 32 *Let $Y \sim \mathcal{G} \in \mathbb{G}_L$, where $\mathbb{G}_L$ is a family of distributions such that $Z_i = \frac{(Y_i - b_i)}{x_i}$ satisfy $E_{\mathcal{G}} Z_i = \mu$, $Var_{\mathcal{G}} Z_i = \sigma^2$, and $Corr_{\mathcal{G}}(Z_i, Z_j) = \rho$ for all $i \neq j$ for some (unknown) $(\mu, \sigma, \rho) \in \Theta$, a parameter space which contains at least two distinct values of μ, and let x_i, b_i be known. Let $s \subset \mathcal{N}$ be fixed and $|s| = n$, and consider the linear predictor*

$$t^* = t^*(s, Y_s) = \frac{n}{N}\bar{Y}_s + \left(1 - \frac{n}{N}\right)(\bar{Z}_s \bar{x}_{s^c} + \bar{b}_{s^c}),$$

where $\bar{Y}_s = \frac{1}{n}\sum_{i\in s} Y_i$, $\bar{Z}_s = \frac{1}{n}\sum_{i\in s} Z_i$, s^c *is the complement of s in $\mathcal{N}$,* $\bar{x}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} x_i$, *and* $\bar{b}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} b_i$. *Then t^* is $\mathcal{G}$-unbiased for any $\mathcal{G} \in \mathbb{G}_L$, and*

$$E_\mathcal{G}(t^*(s, Y_s) - \bar{Y})^2 \leq E_\mathcal{G}(t(s, Y_s) - \bar{Y})^2$$

for any linear predictor t of $\bar{Y}$ that is $\mathcal{G}$-unbiased for all $\mathcal{G} \in \mathbb{G}_L$.

Theorem 33 *Under the conditions of Theorem 32 let now S be a random sample satisfying $S \sim \mathcal{P}$ where $\mathcal{P}$ is any ignorable design. Then*

$$E_{\mathcal{G},\mathcal{P}}(t^*(S, Y_S) - \bar{Y})^2 \leq E_{\mathcal{G},\mathcal{P}}(t(S, Y_S) - \bar{Y})^2$$

for any linear predictor t of the population mean $\bar{Y}$, that is $\mathcal{G}$-unbiased for all $\mathcal{G} \in \mathbb{G}_L$.

Proof of Theorem 33 This follows from the inequality of Theorem 32 by taking $\mathcal{P}$ expectation and exchanging the order of expectations as explained above for ignorable designs. ■

Proof of Theorem 32 The proof is almost the same as in Hedayat and Sinha (1991). We can express any linear predictor in the form

$$t(s, Y_s) = \frac{n}{N}\bar{Y}_s + (1 - \frac{n}{N})\hat{t}(s, Y_s), \text{ where } \hat{t}(s, Y_s) = \sum_{i\in s} c_i Y_i + d.$$

We have $\bar{Y} = \frac{n}{N}\bar{Y}_s + (1 - \frac{n}{N})\bar{Y}_{s^c}$, where $\bar{Y}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} Y_i$, and therefore t is $\mathcal{G}$-unbiased if and only if $E_\mathcal{G}(\hat{t} - \bar{Y}_{s^c}) = 0$, which is equivalent to $\sum_{i\in s} c_i(x_i\mu + b_i) + d = \bar{x}_{s^c}\mu + \bar{b}_{s^c}$. The latter equality holds for two distinct values of μ if and only if

$$\sum_{i\in s} c_i b_i + d = \bar{b}_{s^c}, \text{ and } \sum_{i\in s} c_i x_i = \bar{x}_{s^c}. \tag{36}$$

It suffices to minimize

$$E_\mathcal{G}(\hat{t} - \bar{Y}_{s^c})^2 = Var_\mathcal{G}(\hat{t} - \bar{Y}_{s^c}) = Var_\mathcal{G}\hat{t} + Var_\mathcal{G}\bar{Y}_{s^c} - 2\,Cov_\mathcal{G}(\hat{t}, \bar{Y}_{s^c}).$$

Using (36), it is easy to calculate that $Cov_\mathcal{G}(t, \bar{Y}_{s^c}) = \rho\sigma^2\bar{x}_{s^c}^2$. Therefore, the above minimization is achieved by finding $\hat{t}$ satisfying (36), and having a minimal variance. A straightforward expansion of the variance and (36) lead to

$$Var_\mathcal{G}\hat{t} = \sigma^2[\rho(\sum_{i\in s} c_i x_i)^2 + (1 - \rho)\sum_{i\in s} c_i^2 x_i^2] = \sigma^2[\rho\bar{x}_{s^c}^2 + (1 - \rho)\sum_{i\in s} c_i^2 x_i^2].$$

We can now use the Lagrange method to minimize $\sum_{i \in s} c_i^2 x_i^2$ subject to the constraint $\sum_{i \in s} c_i x_i = \bar{x}_{s^c}$ from (36). We readily obtain the solution $c_i = \bar{x}_{s^c}/nx_i$. From (36) we can now obtain d, and putting it all together with some simple calculations, the result follows. ∎

It is now possible to write an explicit expression of $E_{\mathcal{G}}(t^*(s, Y_s) - \bar{Y})^2$ for any set s, and minimize over s of a given size, thus obtaining an efficient purposive (nonrandom) sample. Such considerations led Royall (1970b) to advocate purposive rather than random sample selection. This approach, and the concept of $\mathcal{G}$-unbiasedness depend on the superpopulation model, unlike man-made randomness and $\mathcal{P}$-unbiasedness where the statistician controls the randomization procedure. The efficiency of purposive designs constructed in the above manner is sensitive to the choice of the prior or superpopulation model and therefore robustness issues arise; see, e.g., Scott, Brewer, and Ho (1978), Hansen, Madow, and Tepping (1983), and references therein. See also Valliant, Dorfman, and Royall (2000), Mukhopadhyay (1998), and Chaudhuri and Stenger (1992) for further discussion and references on the issues arising here, and in other parts of this chapter.

7. Beyond simple random sampling

We have so far concentrated on relatively simple models, and for many results (but not all) on simple sampling designs, with emphasis on (conditional) simple random sampling. We now discuss a few examples of results on various well-known sampling designs, and more general models. Only parts of the results are proved, and other parts are explained or stated without a proof. Here, as in the whole chapter, the results given constitute a sample and certainly not a survey. In all examples below only quadratic loss and the corresponding MSE are considered.

7.1. pps cluster sampling

Related results to Proposition 12, but for **cluster sampling**, with clusters of different sizes and when the estimated parameter is a weighted average (by cluster size) of the cluster means, were given by Scott and Smith (1975), and Scott (1977). They consider ***Bernoulli sampling***, that is, sampling n clusters with replacement, where unit i is drawn with probability p_i in each draw, and in particular the case of ***probability proportional to size (pps) sampling***, where p_i are proportional to cluster size. They show that under certain conditions, the pps strategy minimizes $\sup_y \text{MSE}$ for the pps-Horvitz-Thompson estimator in the class of Bernoulli designs with expected sample size n. When the conditions are relaxed, approximate minimaxity is derived.

7.2. Approximate minimax and the Rao-Hartley-Cochran strategy

We now describe results of Cheng and Li (1983,1987) which extend the results of Section 7.1. Further references can be found in these papers. Consider a population $\mathcal{Y} = (y_1, \ldots, y_N)$ satisfying $y_i = \theta x_i + \varepsilon_i$, $i = 1, \ldots, N$ where $\varepsilon_i = \delta_i g(x_i)$ are nonrandom errors, the x_i's and g are known, and $\delta = (\delta_1, \ldots, \delta_N)$ belongs to some known set L, and $\theta \in \Theta$, some suitable parameter space, is an unknown nuisance parameter.

Given a sample S, a ***linear estimator*** is of the form $t(S, Y) = \sum_{i \in S} r_{si} y_i$, that is, a linear combination of the observations with weights that may depend on S. Let r_s^t (r_s) denotes the row (column) vector $r_s^t = (r_{s1}, \ldots, r_{sN})$, where for $i \notin S$ we set $r_{si} = 0$. For $\mathbf{R} = \{r_s : S \in 2^N\}$ we set $t_{\mathbf{R}}(S, Y) = \sum_{i \in S} r_{si} y_i = r_s^t Y$. A strategy consists of a pair $(\mathcal{P}, t_{\mathbf{R}}(S, Y))$. Our goal is to estimate the population mean $\bar{\mathcal{Y}} = \sum_{i=1}^{N} y_i/N$ using the auxiliary information, and we look for a strategy that minimizes (approximately) the ***risk***

$$\sup_{\theta \in \Theta, \delta \in L} \sum_S \mathcal{P}(S)(\sum_{i \in S} r_{si} y_i - \bar{\mathcal{Y}})^2.$$

Set $\mathbf{x}^t = (x_1, \ldots, x_N)$, $\mathbf{1}$ an N-vector of 1's, $\bar{X} = \sum_{i=1}^{N} x_i/N$, and let G be the $N \times N$ diagonal matrix $G = diag(g(x_1), \ldots, g(x_N))$. We have

$$\sum_S \mathcal{P}(S)(\sum_{i \in S} r_{si} y_i - \sum_{i=1}^{N} y_i/N)^2 = \sum_S \mathcal{P}(S)(\theta\mathbf{x} + G\delta)^t(r_s - \mathbf{1}/N)(r_s - \mathbf{1}/N)^t(\theta\mathbf{x} + G\delta),$$

and it is easy to see that if Θ is unbounded then the risk is bounded if and only if $(r_s - \mathbf{1}/N)^t\mathbf{x} = 0$. If we restrict our choice to such r_s's, then we guarantee that $\sum_{i \in S} r_{si} x_i = \bar{X}$, so that the linear coefficients are **calibrated** for $\bar{X}$. Such a strategy is called ***representative***.

In order to describe one of the results from Cheng and Li (1983) we now define the Rao-Hartley-Cochran (RHC) strategy. In order to obtain a sample of size n, divide the population at random (all partition having equal probabilities) into n groups of predetermined sizes N_j such that $\sum_{j=1}^{n} N_j = N$. Let X_j be the sum of the x_i's in group j. Draws one element from each group, so that if the ith unit is in the jth group, it is drawn with probability x_i/X_j, and denote its y-value by y_j and its x-value by x_j. The Rao-Hartley-Cochran estimator for the population mean $\mathcal{Y}$ is then $t_{RHC} = \sum_{j=1}^{n} X_j y_j/Nx_j$.

Note that this ($\mathcal{P}$-unbiased) estimator is not of the type $t(S, \mathcal{Y}_S)$ considered in this chapter, which are functions of the sampled set S and the corresponding y-values. The quantities X_j are random since they depend on the random partition, with distribution

depending on the data $D[S, \mathcal{Y}_S]$ (actually, on S). Hence it is a **randomized estimator**.[7] With convex loss we could replace X_j by $\mathrm{E}(X_j|S)$ and by Jensen's inequality the risk is reduced (see Remark 4). This calculation is usually complex and therefore it is avoided.

For suitable L and g, and under assumptions relating n, N, and the x_i's which require that n largest x_i's are not too large, Cheng and Li (1983) show that the risk of the RHC strategy is bounded by $1+\varepsilon$ times the maximal risk $\sup_{\theta \in \Theta, \delta \in L} \sum_S \mathcal{P}(S)(\sum_{i \in S} r_{si} y_i - \bar{\mathcal{Y}})^2$, where an explicit bound on ε in terms of the x_i's is given. Thus the RHC strategy is approximate minimax. The details will not be given here.

Cheng and Li (1987) show interesting relation between models as above and **superpopulation** models where the ε_i's are random variables. Such a superpopulation model is considered in Section 7.3.

7.3. Minimax linear esimation in a superpopulation model

Our next discussion concerns a **superpopulation** model that is closely related to the one given in Theorem 32. The results stated here are from Stenger (2002).

Consider a population $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G}$ (see the notation in Section 6.1) generated according to the superpopulation model $Y_i = \theta x_i + \varepsilon_i$, where ε_i are **random variables** satisfying $\mathrm{E}\varepsilon_i = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \gamma u_{ij}$, the x_i's are known, $i, j \in \mathcal{N}$, γ is unknown and the u_{ij}'s are discussed below. Our goal is to estimate the parameter θ on the basis of a sample of size n. Writing $Z_i := Y_i/x_i = \theta + \varepsilon_i/x_i$ we see the similarity to the model of Theorem 32, where now we allow a more general covariance structure.

Parts of the discussion that follows are similar to that of Section 7.2; however, here we are dealing with a superpopulation model. Given a sample S, a ***linear estimator*** is of the form $t(S, Y) = \sum_{i \in S} r_{si} Y_i$, that is, a linear combination of the observations with weights that may depend on S. For $\mathbf{R} = \{r_s : S \in 2^N\}$ we set $t_{\mathbf{R}}(S, Y) = \sum_{i \in S} r_{si} Y_i = r_s^t \mathbf{Y}$, where r_s^t denotes the row vector $(r_{s1}, \ldots, r_{sN})$, and for $i \notin S$ we set $r_{si} = 0$.

Let U denote the $N \times N$ matrix with entries u_{ij}. We assume that $U \in B$, some (known) class of positive definite matrices. For S fixed, the usual MSE decomposition to variance and bias squared yields

$$E_{\mathcal{G}}(\sum_{i \in S} r_{si} Y_i - \theta)^2 = \gamma r_s^t U r_s + \theta^2 (\sum_{i \in S} r_{si} x_i - 1). \tag{37}$$

A strategy consists of a pair $(\mathcal{P}, t_{\mathbf{R}}(S, Y))$. We have

$$E_{\mathcal{G}, \mathcal{P}}(\sum_{i \in S} r_{si} Y_i - \theta)^2 = \sum_S \mathcal{P}(S) E_{\mathcal{G}}(\sum_{i \in S} r_{si} Y_i - \theta)^2,$$

---

[7]Since the probabilities determining the randomization depend on S, t_{RHC} is a *behavioral estimator*.

and here we define a minimax strategy as the strategy $(\mathcal{P}, t_{\mathbf{R}})$ minimizing

$$\sup_{\theta \in \mathbb{R}, U \in B} E_{\mathcal{G},\mathcal{P}} (\sum_{i \in S} r_{si} Y_i - \theta)^2. \tag{38}$$

In view of (37) the expression under the sup is arbitrarily large for large values of θ, unless we set $\sum_{i \in S} r_{si} x_i - 1 = 0$ for all S in the support of $\mathcal{P}$. This condition is equivalent to $E_{\mathcal{G}} \sum_{i \in S} r_{si} Y_i = \theta$ for such S, and hence the same holds for $E_{\mathcal{G},\mathcal{P}}$, and our estimators are unbiased (see Section 6.3). If U is known, that is, $|B| = 1$, the problem reduces to finding a set S that minimizes $\inf_{r_s} \{ r_s^t U r_s : \sum_{i \in S} r_{si} x_i - 1 = 0 \}$, and we conclude that the minimax strategy is degenerate, concentrated at a minimizing set S. Thus for this problem, random sampling is not required. Clearly the minimizing S depends on the x_i's and U. Degenerate designs are called *purposive sampling.*

Next consider $|B| > 1$, and suppose that the covariance matrix U is in the set of diagonal matrices $B = \{ W = diag(w_1, \ldots, w_N) : w_i > 0 \, \forall i, \sum \beta_i w_i \leq 1 \}$ for some given $\beta_1, \ldots, \beta_N > 0$. Since the matrices in B are diagonal, the ε_i's are uncorrelated. Assume $\alpha_i := n\beta_i x_i^2 / \sum_{i=1}^{N} \beta_i x_i^2 \leq 1$ for all $i \in \mathcal{N}$. Stenger's (2002) result is

Theorem 34 *A minimax strategy $(\mathcal{P}_0, t_{\mathbf{R}_0}(S, Y))$, minimizing (38) among strategies consisting of size n designs and linear estimators, is given by any size n design $\mathcal{P}_0$ having inclusion probabilities $\alpha_i = n\beta_i x_i^2 / \sum_{i=1}^{N} \beta_i x_i^2$, and $t_{\mathbf{R}_0} = \frac{1}{n} \sum_{i \in S} Y_i / x_i$.*

Proof By (37) and the discussion following (38), the problem of finding the strategy $(\mathcal{P}, t_{\mathbf{R}})$ minimizing (38), that is, the minimax strategy, is equivalent to minimizing

$$\sup_{W \in B} \sum_{S} r_s^t W r_s \mathcal{P}(S) \quad \text{subject to} \quad r_s^t x - 1 = 0. \tag{39}$$

For a given S, W, and $x = (x_1, \ldots, x_N)$, let $\hat{\theta}_S(W)$ be the linear estimator $\sum_{i \in S} r_{si} Y_i$ derived by minimizing $r_s^t W r_s$ subject to $r_s^t x - 1 = \sum_{i \in S} r_{si} x_i - 1 = 0$. Using Lagrange multipliers we obtain $r_{si} = \frac{x_i / w_i}{\sum_{i \in S} x_i^2 / w_i}$ for $i \in S$, and therefore $\hat{\theta}_S(W) = \frac{\sum_{i \in S} x_i Y_i / w_i}{\sum_{i \in S} x_i^2 / w_i}$. It is easy to see that for any $\mathcal{P}$ the same vectors r_s also minimize $\sum_S r_s^t W r_s \mathcal{P}(S)$ subject to the condition $r_s^t x - 1 = 0$ holding for all S.

By compactness, the sup in (39) is attained at some $V = diag(v_1, \ldots, v_N) \in B$, and therefore the vectors r_s minimizing (39) must satisfy $r_{si} = r_{si}(V) = \frac{x_i / v_i}{\sum_{i \in S} x_i^2 / v_i}$ for $i \in S$. Note that for this $r_s = r_s(V)$ we have $\gamma r_s^t W r_s = \text{Var}_W \hat{\theta}_S(V)$, where Var_W is the variance with respect to the model $\mathcal{G}$ when W is the true covariance matrix. It follows that finding

the minimax strategy is equivalent to finding a design $\mathcal{P}$ and $V \in B$ minimizing

$$\sup_{W \in B} \sum_S Var_W \hat{\theta}_S(V) \mathcal{P}(S). \tag{40}$$

Let $V_0 = diag(v_1, \ldots, v_N)$ where $v_i = x_i^2 / \sum_{i=1}^N \beta_i x_i^2$. Then for $|S| = n$, $\hat{\theta}_S(V_0) = \frac{1}{n} \sum_{i \in S} Y_i / x_i$, $Var_U \hat{\theta}_S(V_0) = \frac{\gamma}{n^2} \sum_{i \in S} u_{ii} / x_i^2$, and $Var_{V_0} \hat{\theta}_S(V_0) = \frac{\gamma}{n}(1/\sum_{i=1}^N \beta_i x_i^2)$, which is independent of S; this will turn out to be useful in (41) below.

Let $\mathcal{P}_0$ be any design with inclusion probabilities $\alpha_i = n\beta_i x_i^2 / \sum_{i=1}^N \beta_i x_i^2$. See Chaudhuri and Vos (1988, Part B) for a survey of methods for construction of such designs. Since $\alpha_i = \sum_{S: S \ni i} \mathcal{P}(S)$, we have for all $V, U \in B$ and any design $\mathcal{P}$

$$\sum_S Var_U \hat{\theta}_S(V_0) \mathcal{P}_0(S) = \frac{\gamma}{n^2} \sum_{i=1}^N \alpha_i u_{ii} / x_i^2 = \frac{\gamma}{n} \sum_{i=1}^N \beta_i u_{ii} / \sum_{i=1}^N \beta_i x_i^2$$

$$\leq \frac{\gamma}{n}(1/\sum_{i=1}^N \beta_i x_i^2) = Var_{V_0} \hat{\theta}_S(V_0) = \sum_S Var_{V_0} \hat{\theta}_S(V_0) \mathcal{P}(S) \tag{41}$$

$$\leq \sum_S Var_{V_0} \hat{\theta}_S(V) \mathcal{P}(S) \leq \sup_{W \in B} \sum_S Var_W \hat{\theta}_S(V) \mathcal{P}(S),$$

where the first inequality holds because $U \in B$, and the second because $V = V_0$ minimizes $Var_{V_0} \hat{\theta}_S(V)$ by definition of $\hat{\theta}_S(W)$. It follows that for any $V \in B$ and any design $\mathcal{P}$

$$\sup_{W \in B} \sum_S Var_W \hat{\theta}_S(V_0) \mathcal{P}_0(S) \leq \sup_{W \in B} \sum_S Var_W \hat{\theta}_S(V) \mathcal{P}(S), \tag{42}$$

and the strategy $(\mathcal{P}_0, \mathbf{r}(V_0))$ minimizes the expression in (40), and hence it is the minimax strategy in the sense defined in (38). ∎

Unlike the result of Theorem 32, which concerns estimation or prediction of the population mean $\bar{Y}$, the problem of estimating the regression parameter θ discusses above and the sample selection based on the x_i's may be viewed as belonging to the area of *optimal regression design* rather than sampling. Note in particular that even if the whole population $(Y_1, \ldots, Y_N)$ is observed, the parameter θ is not determined. Stenger (2002) discusses also the problem of predicting $\bar{Y}$, under the same regression model, and proves existence of minimax strategies. Again, purposive sampling suffices when $|B| = 1$, and random sampling is required for $|B| > 1$.

Returning to the problem of estimating the population mean under a **superpopulation** model, we now discuss the seminal work of Aggarwal (1959, 1966). The population $Y = (Y_1, \ldots, Y_N)$ is distributed according to $\mathcal{G} \in \mathbb{H}$, where $\mathbb{H}$ is the class of distributions

$\mathcal{G}$ that are concentrated on a hyperplane in $\mathbb{R}^N$ of the form $Y_1 + \ldots Y_N = $ constant, say $N\mu_{\mathcal{G}}$, and subject to

$$E_{\mathcal{G}} \sum_{i=1}^{N} (Y_i - \mu_{\mathcal{G}})^2 \leq M \tag{43}$$

for some $M > 0$. Setting $E_{\mathcal{G}} Y_i = \mu_{\mathcal{G},i}$, and $\text{Var}_{\mathcal{G}} Y_i = \sigma_{\mathcal{G},i}^2$, we can express that latter condition as $\sum_{i=1}^{N} [\sigma_{\mathcal{G},i}^2 + (\mu_{\mathcal{G},i} - \mu_{\mathcal{G}})^2] \leq M$. The goal is to estimate the population mean $\bar{Y}$ which here equals $\mu_{\mathcal{G}}$. Under $\mathcal{P}_s$, simple random sampling of n observations, consider the problem of finding the minimax estimator, that is, the estimator t minimizing the risk $\sup_{\mathcal{G} \in \mathbb{H}} E_{\mathcal{P}_s, \mathcal{G}} (t(S, Y) - \bar{Y})^2$. Aggarwal (1959) uses Bayesian calculations (see Section 3.5) to show that the minimax estimator is the sample mean.

If the population is divided into given **strata**, and simple random sampling with a given sample size is carried out in each stratum, and if in each stratum a superpopulation model of the above kind holds (with the bound M_i instead of M of (43) for the ith stratum), then the usual weighted sum of the strata means is shown to be minimax. For a statistician who can choose the sample sizes, and the cost of sampling is added to the above risk, Aggarwal (1959) provides the minimax strategy[8], consisting of the same weighted mean, and where naturally the sample sizes in the strata depend on the bounds M_i, and the cost of sampling in each stratum.

Aggarwal (1966) provides similar results for ***two-stage sampling***. Now the population is divided into given subgroups called ***primary units*** (or **clusters**). A simple random sample of primary units (clusters) is selected in the first stage (whereas in stratified sampling all strata are sampled), and a second-stage simple random sampling is carried out in each of the selected clusters. The superpopulation model constrains the cluster means to be on a hyperplane, as well as the Y's within each cluster, with conditions similar to (43) within and between clusters, and suitable bounds replacing M. With weights computed in terms of these bounds and the sample sizes, a weighted average of the sample means in the sampled clusters is shown to be minimax for given sample sizes. A minimax allocations of sample sizes that depends on the bounds and the sampling costs is also given, which together with the above estimator comprise a minimax strategy.[8]

8. List of main notations

$\mathcal{Y} = (y_1, \ldots, y_N)$, a finite population of size N. $\mathcal{N} = \{1, \ldots, N\}$. $S \subseteq \mathcal{N}$, a sample.
$\mathcal{Y}_S$ - the multiset containing all y_i-values arising from distinct labels $i \in S$.

---

[8]Among strategies based on simple random sampling.

$Y = (Y_1, \ldots, Y_N)$, a random finite population of size N under a superpopulation model.
$\mathcal{G}$ - distribution of Y (prior or superpopulation model). $\mathbb{G}$ - class of exchangeable priors.
Y_S - the multiset containing all Y_i-values arising from distinct labels $i \in S$.
$\bar{\mathcal{Y}} = \frac{1}{N} \sum_{i=1}^{N} y_i$, the population mean. $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$, the population mean under a superpopulation model.
$\bar{\mathcal{Y}}_S = \frac{1}{n} \sum_{i \in S} y_i$, the sample mean, where $n = |S|$. $\bar{Y}_S = \frac{1}{n} \sum_{i \in S} Y_i$, the sample mean under a superpopulation model.
$D = D[S, \mathcal{Y}] = \{(i, y_i) : i \in S\}$, the data.
$t = t(D) = t(\{(i, y_i) : i \in S\})$ - an estimator. We also use $t(D[(S, \mathcal{Y}_S)])$, $t(S, \mathcal{Y}_S)$, or $t(S, \mathcal{Y})$. Under a superpopulation model we use $t(\{(i, Y_i) : i \in S\})$ or $t(D[(S, Y_S)])$, etc.
$\mathcal{P}$ - a sampling design (probability over subsets S of $\mathcal{N}$). $\alpha_i = \mathcal{P}(\{i \in S\})$, inclusion probabilities.
SRS $= \mathcal{P}_s$ - simple random sampling without replacement, also SRS.
$t_{HT} = \sum_{i \in S} y_i / \alpha_i$ - the Horvitz-Thompson estimator.
$L(t, \mathcal{Y})$ - loss when the estimator takes the value t.
$R(\mathcal{P}, t; \mathcal{Y}) := E_{\mathcal{P}} L(t, \mathcal{Y}) = \sum_S \mathcal{P}(S) L(t(D[S, \mathcal{Y}_S]), \mathcal{Y})$ - risk.

REFERENCES

1. Aggarwal., OM. P., (1959). Bayes and minimax procedures in sampling from finite and infinite populations-I. Ann. Math. Statist. 30, 206-218.
2. Aggarwal., OM. P., (1966). Bayes and minimax procedures for estimating the arithmetic mean of a population with two-stage sampling. Ann. Math. Statist. 37, 1186-95.
3. Basu, D., (1958). On sampling with and without replacement. Sankhya 20, 287-294.
4. Basu, D., (1971). An Essay on the Logical Foundations of Survey Sampling, Part 1 (with discussion), in Foundations of Statistical Inference, eds. V. P. Godambe and D. A. Sprott. Toronto: Holt, Rinehart and Winston, pp. 203-242.
5. Bickel, P. J., and Lehmann, E. L. (1981). A minimax property of the sample mean in

finite populations. Ann. Statist. 9, 1119-1122.

6. Blackwell, D., and Girshick, M. A., (1954). Theory of Games and Statistical Decisions. New York: Wiley.

7. Brewer, K. R. W., (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. Austral. J. Statist. 5, 93-105.

8. Cassel, C. M., Särndal, C. E., and Wretman, J. H., (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

9. Chaudhuri, A. and Stenger, H., (1992). Survey sampling: theory and methods. Marcel Dakker Inc NY.

10. Chaudhuri, A. and Vos, J., (1988). Unified Theory and Strategies of Survey Sampling. Amsterdam: North-Holland.

11. Cheng, C. S., and Li, K. C., (1983). A minimax approach to sample surveys. Ann. Statist. 11, 552-563.

12. Cheng, C. S., and Li, K. C., (1987). Optimality criteria in survey sampling. Biometrika 74, 337-345.

13. Cochran, W. G., (1963). Sampling Techniques. New York: Wiley.

14. Ferguson, T. S., (1967). Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press, Inc.

15. Gabler, S., (1990). Minimax Solutions in Sampling from Finite Populations. Springer-Verlag, Berlin.

16. Gabler, S., and Stenger, H., (2000). Minimax strategies in survey sampling. Journal of Statistical Planning and Inference 90, 305-321.

17. Godambe, V. M., (1955). A unified theory of sampling from finite populations. J. R. Statist. Soc. B 17, 269-278.

18. Godambe, V. M., and Joshi, V. P., (1965). Admissibility and Bayes estimation in sampling finite populations I. Ann. Math. Statist. 36, 1707-1722.

19. Hansen, M. H., Madow, W.G, and Tepping, B. J., (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Amer. Statist. Assoc. 78, 776-793.

20. Hedayat, A. S., and Sinha, B. K., (1991). Design and Inference in Finite Population Sampling. Wiley, New York.

21. Helland, I. S., (1995). Simple counterexamples against the conditionality principle. Amer. Statistician 49, 351-356.

22. Hodges, J. L. Jr., and Lehmann, E. L., (1982). Minimax estimation in simple random sampling, in eds. G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh. Statistics and

Probability: Essays in Honor of C. R. Rao. North-Holland, Amsterdam, pp. 325-327.

23. Joshi, V. M., (1965). Admissibility and Bayes estimation in sampling finite populations III. Ann. Math. Statist. 36, 1730-1742.

24. Joshi, V. M., (1966). Admissibility and Bayes estimation in sampling finite populations IV. Ann. Math. Statist. 37, 1658-1670.

25. Kiefer, J., (1957). Invariance, minimax sequential estimation, and continuous time processes. Ann. Math. Statist. 28, 573-601.

26. Lahiri, D. B., (1951). A method of sample selection providing unbiased ratio estimates. Bull. International Statist. Inst., 33, 133-40.

27. Lehmann, E. L., and Casella, G., (1998). Theory of Point Estimation, 2nd ed. Springer-Verlag, New York.

28. Mukhopadhyay, P., (1998). Theory and Methods of Survey Sampling. Prentice-Hall of India, New Delhi.

29. Ramakrishnan, M. K., (1973). An alternative proof of the admissibility of the Horvitz-Thompson estimator. Ann. Statist. 1, 577-579.

30. Rao, J. N. K., (1971). Some thoughts on the foundations of survey sampling. J. Indian Soc. Agricultural Statist. 23, 69-82

31. Rao, J. N. K., Hartley, H. O., and Cochran, W. G., (1962). On a simple procedure of unequal probability sampling without replacement. J. Royal Statist. Soc. B 24, 482-491.

32. Rao, J. N. K., and Vijayan, K., (1977). On estimating the variance in sampling with probability proportional to aggregate size. J. Amer. Statist. Assoc. 72, 579-584.

33. Royall, R., (1968). An old approach to finite population sampling theory. J. Amer. Statist. Assoc. 63, 1269-1279.

34. Royall, R. M., (1970a). Finite population sampling - On labels in estimation. Ann. Math. Statist. 41, 1774-1779.

35. Royall, R. M., (1970b). On finite population sampling theory under certain linear regression models. Biometrika 57, 377-387.

36. Rubinstein, A., (1991). Comments on the interpretation of game theory. Econometrica 59, 909-924.

37. Särndal, C., Swensson, B., and Wretman, J., (1992). Model assisted survey sampling. New York: Springer-Verlag.

38. Scott, A. J., (1975). On admissibility and uniform admissibility in finite population sampling. Ann. Statist. 3, 489-491.

39. Scott, A. J., (1977). On the problem of randomization in survey sampling. Sankhya Ser. 39, 1-9.

40. Scott, A. J., and Smith, T. M. F., (1975). Minimax designs for sample surveys. Biometrika 62, 353-357.

41. Scott, A. J., Brewer, K. R. W., and Ho, E. W. H., (1978). Finite population sampling and robust estimation. J. Amer. Statist. Assoc. 73, 359-361.

42. Smith, T.M.F., (1976). The foundations of survey sampling: a review. J. Roy. Statist. Soc. Ser. A 139, 183-204.

43. Stenger, H., (1979). A minimax approach to randomization and estimation in survey sampling. Ann. Statist. 7, 395-399.

44. Stenger, H., (2002). Regression analysis and random sampling. J. Statist. Plan. Infer. 102, 169178.

45. Stenger, H., Gabler, S., and Schmidt, J., (2002). Survey sampling: a linear game. Statistics and Decisions 20, 279-295.

46. Thompson, S.K., and Seber, G. A. F., (1996). Adaptive Sampling. J. Wiley & Sons : New York.

47. Valliant, R., Dorfman, A. H., and Royall, R. M., (2000). Finite population sampling and inference: a prediction approach. Wiley New York.

Table of contents

Key words

sampling design, noninformative design, adaptive design conditional simple random sampling, parameter space, symmetric parameter space, inclusion probability, Horvitz-Thompson estimator, strategy, minimax strategy, unbiased strategy, loss function, quadratic loss, convex loss, symmetric parameter, estimator, symmetric estimator, $\mathcal{P}$-unbiased estimator, randomized estimator, Bayes estimator, generalized difference estimator, admissible estimator, superpopulation model, risk, predictor, $\mathcal{G}$-unbiased predictor, linear predictor, cluster sampling, Bernoulli sampling, probability proportional to size (pps) sampling.