Proceedings of Symposia in Applied Mathematics
Volume 23

MODERN STATISTICS: Methods and Applications

Edited by
ROBERT V. HOGG

PROCEEDINGS OF SYMPOSIA
IN APPLIED MATHEMATICS
Volume XXIII

MODERN STATISTICS:
METHODS AND APPLICATIONS

AMERICAN MATHEMATICAL SOCIETY
PROVIDENCE, RHODE ISLAND
1980

LECTURE NOTES PREPARED FOR THE

AMERICAN MATHEMATICAL SOCIETY SHORT COURSE

MODERN STATISTICS: METHODS AND APPLICATIONS

HELD IN SAN ANTONIO, TEXAS
JANUARY 7–8, 1980

EDITED BY
ROBERT V. HOGG

The AMS Short Course Series is sponsored by the Society's Committee on Employment and Educational Policy (CEEP). The Series is under the direction of the Short Course Advisory Subcommittee of CEEP.

Library of Congress Cataloging in Publication Data

American Mathematical Society Short Course on Modern Statistics: Methods and Applications, San Antonio, 1980.

Modern statistics, methods and applications.

(Proceedings of symposia in applied mathematics; v. 23)

"Lecture notes prepared for the American Mathematical Society Short Course [on] Modern Statistics: Methods and Applications, held in San Antonio, Texas, January 7–8, 1980."

Includes bibliographies.

1. Mathematical statistics—Congresses. I. Hogg, Robert V. II. American Mathematical Society. III. Title. IV. Series.

QA276.A1A45 1980 001.4'22 80-16093

ISBN 0-8218-0023-X

1980 *Mathematics Subject Classification.* Primary 62D05, 68G10, 62G05, 62M10, 62F35.

PREFACE

This volume contains the lecture notes prepared by the speakers for the American Mathematical Society Short Course on *Modern Statistics: Methods and Applications* given in San Antonio, Texas, on January 7–8, 1980.

We were very pleased with the substantial attendance at the course. The skills of the lecturers and the enthusiasm of the participants encouraged the AMS Committee on Short Courses to request that these notes be published. We are indebted to our colleagues for this support and the AMS office for the cooperation in publishing these proceedings.

Of course, the choice of topics from a field as large as Statistics is a difficult one. However, I did want to avoid any substantial overlap with the highly successful short course on statistics held three years earlier in St. Louis, January, 1977. Therefore it seemed very natural to begin with one important topic that is sometimes overlooked in an introductory course, particularly one in mathematical statistics. Yet this topic is one through which the general public most often hears about statistics, namely, survey sampling. Wayne Fuller spoke on "Samples and Surveys", noting the operations necessary in conducting a survey of a human population. In his article, he explains the construction of a probability sample design and the corresponding optimal estimators.

The more general problem of the design and analysis of an experiment was covered by Peter John in his "Analysis of Variance". These techniques have, for years, been extremely important in applications and have also motivated a large amount of statistical research. It is clear that even in an elementary design the experimenter must understand the importance of randomization.

Nonparametric statistical methods have played a major role in modern statistics. Two coordinated talks on that subject were given by Ronald Randles and Thomas Hettmansperger. Randles introduced distribution-free rank tests, such as one by Wilcoxon, and some of their good asymptotic properties. Hettmansperger then explained how these rank tests could be used to obtain point and interval estimates for various parameters, including the regression situation. These resulting

R-estimates are very robust because they are not highly sensitive to reasonable deviations from the underlying assumptions.

The important topic of regression was continued by considering isotonic regression and time series. F. T. Wright showed how to use the method of maximum likelihood to estimate ordered parameters. Then Douglas Martin considered a time sequence of data. After presenting a collection of interesting examples, he discussed appropriate models and their estimates, including robust ones.

While it is impossible to cover all of Statistics in six articles, these and their references should prove useful to those who wish to learn something of the natures of modern statistics. In that regard, I must also call your attention to *Studies in Statistics* that I had the opportunity to edit for Volume 19 of Studies in Mathematics under the sponsorship of the Mathematical Association of America. I hope that this present volume, along with that one, will provide the interested reader a good introduction to modern statistical methods.

Robert V. Hogg
University of Iowa
March, 1980

CONTENTS

Proceedings of Symposia in Applied Mathematics
Volume 23
1980

SAMPLES AND SURVEYS

Wayne A. Fuller[1], Iowa State University

I. Introduction.

The design and execution of a large scale survey is a sizeable research undertaking. We outline the steps in such an operation.

A. Definition of the objectives.
B. Specification of the procedures.
   1. Universe of interest.
   2. Data to be collected and method of collection.
   3. Sample design.
   4. Questionnaire design.
C. Field work.
D. Data processing.
   1. Coding.
   2. Editing.
   3. Estimation and tabulation.
E. Report preparation.

We assume that the objectives of the study require obtaining data from an existing group of elements. The universe is the collection of elements about which statements will be made. In most surveys data are collected on a large number of characteristics. The regular polls (Gallup, etc.) record items such as age, race, sex, place of residence, and political affiliation, in addition to responses on a few questions of current interest.

The reasons for observing a part of the universe (taking a sample) instead of the entire universe (a census) are all practical. First the research budget seldom permits observing every element of the population. A personal interview now costs on the order of $30 to $100 to complete. Also, in certain quality testing situations, the observations are destructive. It is of little use to know that a lot of light bulbs will last an average of

---

1980 Mathematics Subject Classification 62D05.
[1]This research was partly supported by Joint Statistical Agreement JSA 79-10 with the U.S. Bureau of the Census.

200 hours if they have all been burned to establish that fact.

There are other disadvantages of censuses. The first is timing. The data for the 1970 Census of Population were collected beginning in April 1970. The first preliminary raw count reports (for states, counties and municipalities) were available in May through October 1970. The advanced reports become available in the period September 1970 through February 1971. The U.S. summary report was released in January 1972.

More subtle is the problem of quality control in a census. The population census of the United States requires over one quarter million field workers and supervisors. Because fewer interviewers are required for a sample, it should be possible to select better people and to better supervise the field operation.

Once one has decided that a census is impossible, the questions become: What kind of sample? How large a sample?

A sample is a portion of the universe. A random sample (or probability sample) is a sample selected in such a way that the probability of selecting every sample is known.

A simple random sample is a sample of n elements chosen from a population of N elements in such a way that each one of the ${}_{N}C_{n}$ samples has an equal probability of being selected.

A purposive sample (an alternative term is judgmental sample) is any sample that is not a probability sample. Generally speaking, purposive samples are selected to meet certain criteria. The prime example is the political subdivision that has voted for the winner in the last ten elections.

What is the place of the two kinds of sampling? Let us first consider the problem from an empirical point of view. An experiment cited by Jessen (1978, p. 18) compared two methods of sampling a universe of 126 stones. Members of a statistics class were instructed to look at the entire universe and then to select a sample that would represent the average weight of the stones. The sixteen students selected three samples of sizes 1, 2, 5 and 10 and one sample of size 20. Simple random samples of the same sizes were also selected. (126 samples of size one, 30 of size 2, 90 of size 5, 60 of size 10 and 10 of size 20 were selected.)

Table 1. Mean absolute deviation for two types of sample selection.[1]

Type of Sample	Sample Size				
	1	2	5	10	20
Judgment	40.0	44.9	35.3	38.5	31.0
Random	80.6	71.4	41.3	34.1	26.2

[1]From Jessen (1978, p. 18)

The conclusion of the Jessen experiment seems a part of scientific practice. That is, you can expect that your journal article will be accepted if you are working with a very small judgment sample, on the order of 5 or less, but can anticipate difficulties if you submit an article based on a large judgment sample.

II. Simple Random Sampling.

We present a few of the results on simple random sampling. Because of the simplicity of these results, sampling offers an excellent method of introducing a student to statistics. The population of possible samples can be enumerated and the expectation of a random variable can be introduced as the average over the finite number of possible outcomes.

Let the population be composed of N elements. Let the value of the y-characteristic of the elements be denoted by $\{y_i: i=1, 2, \ldots, N\}$. The probability that a particular element appears in a simple random sample is n/N. The probability that a particular pair of elements appears in the sample is $[N(N-1)]^{-1} n(n-1)$. From these basic properties of simple random samples several results are immediate.

RESULT 1. The sample mean is unbiased for the population mean.

RESULT 2. The variance of the sample mean is

$$E\{(\bar{y} - \bar{Y})^2\} = \frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{N-n}{N}\frac{S^2}{n} , \tag{1}$$

where $S^2 = N(N-1)^{-1}\sigma^2$, $\sigma^2 = N^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $\bar{y}$ is the sample mean.

RESULT 3. An unbiased estimator of S^2 is

$$s^2 = (n-1)^{-1}\sum_{i=1}^{n}[y_i - \bar{y}]^2 . \tag{2}$$

Assume that the characteristic y takes on the two values 0 and 1. Let N_1 of the elements be ones and $N-N_1$ of the elements be zeros. If a simple random sample of size n is selected from the N elements, the probability that exactly n_1 of the elements will possess a y-charcteristic of 1 is

$$P\left(\sum_{i=1}^{n} y_i = n_1\right) = \frac{\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N}{n}} \tag{4}$$

Table 2 contains the probabilities for all possible values of (N_1, n_1) for a sample of $n=5$ selected from a population of $N=15$. Two lines have been drawn through the values. The lines are such that the sum of the

probabilities to the right of the right line in every row is less than 0.12 .

The lines enable us to define an interval for each sample outcome such that, for each value of N_1, the probability is greater than 0.88 that the interval will cover the true N_1. The intervals for each n_1 are defined by the horizontal lines in Table 2. The intervals of Table 2 are the classical confidence intervals introduced by Neyman (1934, 1935). In particular, see Neyman (1934, p. 624).

Table 2. Probabilities of sample outcomes for samples of size five selected from fifteen

Number of 1's in Population	Number of 1's in Sample n_1					
N_1	0	1	2	3	4	5
0	1.0000	0	0	0	0	0
1	0.6667	0.3333	0	0	0	0
2	0.4286	0.4762	0.0952	0	0	0
3	0.2637	0.4945	0.2198	0.0220	0	0
4	0.1538	0.4395	0.3297	0.0733	0.0037	0
5	0.0839	0.3497	0.3996	0.1499	0.0166	0.0003
6	0.0420	0.2517	0.4196	0.2398	0.0449	0.0020
7	0.0187	0.1632	0.3916	0.3263	0.0932	0.0070
8	0.0070	0.0932	0.3263	0.3916	0.1632	0.0187
9	0.0020	0.0449	0.2398	0.4196	0.2517	0.0420
10	0.0003	0.0166	0.1499	0.3966	0.3497	0.0839
11	0	0.0037	0.0733	0.3297	0.4395	0.1538
12	0	0	0.0220	0.2198	0.4945	0.2637
13	0	0	0	0.0952	0.4762	0.4286
14	0	0	0	0	0.3333	0.6667
15	0	0	0	0	0	1.0000

Table 3. Possible outcomes for a population with $N_1 = 8$

n_1	Probability	Interval	Statement
0	0.0070	[0, 4]	Wrong
1	0.0932	[1, 7]	Wrong
2	0.3263	[3, 10]	Right
3	0.3916	[5, 12]	Right
4	0.1632	[8, 14]	Right
5	0.0187	[11, 15]	Wrong

The possible outcomes for $N_1=8$ are given in Table 3. The probability of a wrong statement is 0.1189 and the probability of a correct statement is 0.8811.

The confidence interval is particularly forceful when used with random sampling of zero-one characteristics. This is because the model is guaranteed to be correct. The use of randomization in sample selection creates the population of samples for which our statement applies. In random sampling from a finite universe, we know the nature of the sampling distribution because we have created it!

To establish confidence intervals for the mean of a finite population for a characteristic that is not 0-1 we must expand our theoretical constructs. One approach is to assume that the finite population is a random sample from an infinite superpopulation. For example, assume that the finite population is a random sample from a normal population. If we select a random sample from the finite population, we have created two independent random samples, one of size n and one of size $N-n$, from the original population. Then the difference $\bar{y} - \bar{Y}$, where $\bar{y}$ is the mean of the n elements and $\bar{Y}$ is the mean of the N elements, is distributed as a normal random variable with mean zero and variance

$$V\{\bar{y} - \bar{Y}\} = \frac{N-n}{Nn}\sigma^2 .$$

It follows that

$$t = [(Nn)^{-1}(N-n)s^2]^{-\frac{1}{2}}(\bar{y} - \bar{Y})$$

is distributed as Student's t. The denominator of the t statistic is an unbiased estimator of the variance conditional on $(y_1, y_2, \ldots, y_N)$ as well as an unbiased estimator of the unconditional variance. In this argument the distribution of $\bar{y} - \bar{Y}$ is for the population of all possible pairs of samples of size n and $N-n$ selected from the parent population. Perhaps, because samplers have traditionally preferred to think of the finite population as fixed, this theoretical construct seldom appears in sampling texts.

In defining a sequence for a central limit theorem for sampling from a finite universe, one must consider a sequence of samples selected from a sequence of populations. To obtain a limiting normal distribution, the sequence of populations must satisfy certain conditions. Two approaches have been used. One is to specify conditions on the population sequence itself. In this case the distribution is for the population of random samples created by randomization conditional on fixed population values. Madow (1948) and Hájek (1960) give results of this type. The second approach is to assume that the finite population is a sample from an infinite population. We state a result of the second type.

THEOREM. Let $\{\mathcal{U}_t : t=1, 2, \ldots\}$ denote a sequence of finite popula-

tions, where $\mathcal{U}_t$ is a random sample of size N_t, $N_t > N_{t-1}$, selected from an infinite population. Assume the infinite population possesses finite first and second moments. Let simple random samples of size n_t, $n_t > n_{t-1}$ be selected from N_t. Let

$$\lim_{t \to \infty} N_t^{-1} n_t = f, \qquad 0 \leq f < 1 .$$

Then

$$n_t^{\frac{1}{2}} (\bar{y}_t - \bar{Y}_t) \xrightarrow{\mathcal{L}} N(0, (1-f)\sigma^2) ,$$

where $\bar{y}_t$ is the sample mean and $\bar{Y}_t$ is the population mean for population t and σ^2 is the variance of the infinite population.

III. Unequal Probability Sampling.

To use probability ideas in sampling, it is not necessary that each element have an equal probability of entering the sample. To introduce the ideas of unequal probability sampling, consider a population of five elements with characteristics $\{y_1, y_2, y_3, y_4, y_5\}$. Assume that we create ten slips of paper as given in Table 4. Let p_i denote the fraction of the slips that have element i recorded on them. Then the average over the ten slips of the ratios $p_i^{-1} y_i$ is $\Sigma_{i=1}^{5} y_i$. Therefore, if we randomly choose one slip, an unbiased estimator of the total of y is $p_i^{-1} y_i$. If we use replacement sampling, the average of the n values we observe is an unbiased estimator of the population total Y with variance

$$V\{\hat{Y}\} = n^{-1} \sum_{i=1}^{N} p_i (p_i^{-1} y_i - Y)^2 . \tag{8}$$

In nonreplacement sampling, the probabilities of selection π_i are typically specified so that $\Sigma_{i=1}^{N} \pi_i = n$, where n is the number of elements to be included in the sample. With this normalization, the unbiased estimator of the population total is

$$\hat{Y} = \sum_{j=1}^{n} y_j \pi_j^{-1} .$$

The variance of the estimator depends upon the joint probabilities of selection π_{ij};

$$V\{\hat{Y}\} = \sum_{i=1}^{N} \pi_i^{-1} (1-\pi_i) y_i^2 + \sum_{i \neq j}^{N} \pi_i^{-1} \pi_j^{-1} (\pi_{ij} - \pi_i \pi_j) y_i y_j . \tag{9}$$

An unbiased estimator of this quantity is

$$\hat{V}\{\hat{Y}\} = \sum_{i<j}^{n} \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij})[\pi_i^{-1} y_i - \pi_j^{-1} y_j]^2 . \qquad (10)$$

Table 4. Population of slips for unequal probability sampling.

Original Population Element	y-value	Number of Such Slips	$p_i^{-1} y_i$
1	y_1	4	$2.5\ y_1$
1	y_1	4	$2.5\ y_1$
1	y_1	4	$2.5\ y_1$
1	y_1	4	$2.5\ y_1$
2	y_2	3	$(10/3)\ y_2$
2	y_2	3	$(10/3)\ y_2$
2	y_2	3	$(10/3)\ y_2$
3	y_3	1	$10\ y_3$
4	y_4	1	$10\ y_4$
5	y_5	1	$10\ y_5$

IV. Sample Sizes and Sample Frames.

The first question a consulting survey statistician hears from the client is: How many ... do I need? The question, formulated so that an answer is possible, requires considerable information:

(1) A statement of desired closeness for the final answer. For example: "I wish my estimate of the mean of y to be within d units of the true value with probability $1-\alpha$."

(2) An estimate of the variability in the parent population. For example: "The variable y is similar to the variable x which has a variance of σ^2."

Assume the existence of an idealized client that specifies that the estimate of the proportion is to be within 0.02 of the true proportion with probability $1-\alpha$. Some required sample sizes are given in Table 5.

To select a probability sample, it is necessary to create a list, called the sampling frame, such that every element of the universe is associated with at least one item on the list. We consider frames such that each element is associated with one and only one item on the list. The frame may be a physical list such as a list of automobile registrations or it may be a con-

ceptual list such as the list of all possible latitude, longitude coordinates for every point of the land area in the United States.

Table 5. Size of sample required for observed proportion p to be within 0.02 of population proportion P with at least probability $1-\alpha$.

Population Size	True Proportion P = 0.5		P = 0.1
	$1-\alpha = 0.95$	$1-\alpha = 0.99$	$1-\alpha = 0.99$
100	98	99	98
1,000	720	816	610
2,500	1,226	1,559	948
10,000	1,922	2,946	1,304
25,000	2,206	3,571	1,424
50,000	2,306	3,844	1,465
100,000	2,359	3,996	1,482
∞	2,401	4,147	1,493

The construction of a standard sampling frame is the task of constructing a list of primary sampling units such that every element in the population is in exactly one of the primary sampling units. Some primary sampling units may contain no elements and some may contain several. Because so few lists of human and economic populations exist, it is often necessary to create a list of primary sampling units that can be identified in the field operation. One of the most important frames of this type is the area frame.

The area sample is an example of a cluster sample. If any primary sampling unit in the frame contains more than one or less than one observation unit (element) the primary sampling units are called clusters.

Cluster sampling is used for two reasons.

(a) It may be impossible or prohibitively expensive to construct a list of observation units.

(b) For a fixed expenditure, it is often possible to obtain a smaller mean square error for an estimator by observing groups of observation units.

The estimation formulas presented for simple random samples apply for cluster samples with y_i being the total of the y-characteristics of the elements in the primary sampling unit.

VII. Ratio and Regression Estimation.

Methods of estimation employed in survey sampling, beyond the basic mean-

type estimators we have presented, are ratio and regression estimation. Assume that we have available some information about the population. As an example, consider a study of farms. We have very good data on the land in farms. Call the total land in farms X and assume it to be known. Assume, less realistically, that we draw a random sample of farms. Let the y-characteristic be the acres of corn. Then the ratio estimator of the total acres of corn is

$$\hat{Y}_r = \left(\sum_{i=1}^{n} y_i \right) \left(\sum_{i=1}^{n} x_i \right)^{-1} X = \bar{x}^{-1} \bar{y} X , \tag{21}$$

where x_i is the acreage of the i^{th} sample farm. The simple regression estimator of the total acres of corn is

$$\hat{Y}_\ell = N[\bar{y} + b(\bar{X} - \bar{x})] , \tag{22}$$

where $\bar{X} = N^{-1} X$ and b is the usual least squares regression coefficient.

Neither of these estimators is unbiased. Defining a sequence of populations, it is possible to demonstrate that $n^{\frac{1}{2}}(\bar{y}_r - \bar{Y})$ is approximately distributed with mean zero and variance

$$(1 - f)(S_Y^2 - 2 R S_{XY} + R^2 S_X^2) , \tag{23}$$

where $f = N^{-1} n$, $R = \bar{Y}/\bar{X}$, and $\bar{y}_r = N^{-1} \hat{Y}_r$. Similarly, $n^{\frac{1}{2}}(\bar{y}_\ell - \bar{Y})$ is approximately distributed with mean zero and variance

$$(1 - f)(S_Y^2 - B S_{XY}) , \tag{24}$$

where $B = S_X^{-2} S_{XY}$ and $\bar{y}_\ell = N^{-1} \hat{Y}_\ell$.

VIII. Survey Design.

The objective of survey design is to use the available information to create a method of sampling and an estimation rule that yields estimators with desirable properties. Some desirable properties of design-estimator pairs are:

1. Unbiasedness.
2. Accuracy. A measure of the accuracy of an estimator $\hat{\theta}$ of θ is the mean square error.

   $$\text{M.S.E.} = E\{(\hat{\theta} - \theta)^2\}$$

3. Consistency.
4. Scale invariance. The estimator $\hat{\theta}(y)$ is scale invariant if $\hat{\theta}(ky) = k\hat{\theta}(y)$ for all fixed k.
5. Location invariance. The estimator $\hat{\theta}(y)$ is location invariant if

$\theta(k+y) = k+\theta(y)$ for all fixed k.

6. Simplicity.
7. Internal consistency. The estimators $\hat{\theta}_1(y)$, $\hat{\theta}_2(y)$ and $\hat{\theta}_3(y)$ are internally additively consistent for θ_1, θ_2, and θ_3 if

$$\hat{\theta}_1(y) + \hat{\theta}_2(y) = \hat{\theta}_3(y)$$

for $\theta_1 + \theta_2 = \theta_3$.

8. Accurate internal estimator of the variability of the estimator.
9. Robustness. A sample design-estimator pair is robust if departures of the sampled population from that anticipated at the design stage produce small decreases in accuracy.
10. Practicality.

Godambe (1955) pointed out that if we pay attention to the individual values of the population (treat them as fixed constants) there is no probability design that is best for all possible populations. For a particular set of positive y's we can obtain zero variance by making the selection probabilities p_i of Table 5 proportional to y_i.

Godambe's result suggests that we should quantify the prior information associated with the individual elements of the population at the design stage. We use Godambe's formalism to present the design problem.

Let $\mathcal{U} = \{u_i: i=1, 2, \ldots, N\}$ denote the N units of the finite population. Let s denote a subset of the units of the population and let $\mathcal{S}$ be the set of all subsets, s. Let d be an estimator constructed from the units of s. Let p denote a sampling design. The sampling design assigns probabilities, summing to one, to the elements s of $\mathcal{S}$. Let $\mathcal{G}$ denote the prior information about $\mathcal{U}$ available at the time a survey design is to be chosen. Let $\mathcal{P}$ denote the set of all possible designs and $\mathcal{D}$ the set of all possible estimators. Given $\mathcal{G}$ and an optimality criterion, the problem of choosing the optimal $p \subset \mathcal{P}$ and $d \subset \mathcal{D}$ is the problem of survey design.

The class $\mathcal{D}$ is often restricted to the class of linear estimators. A linear estimator can be written as

$$d = \sum_{i=1}^{n} w_i(s)\, y_i \tag{25}$$

where the weights are permitted to be a function of the element identification and of the sample.

Let us consider a simple statistical problem. Let the observations $\{y_i: i=1, 2, \ldots, n\}$ be independently distributed (μ, σ^2) where μ and σ^2 are unknown. That is, $\{y_1, y_2, \ldots, y_n\}$ is a random sample selected from a parent distribution with mean μ and variance σ^2. Then $\bar{y}$ is the

best linear unbiased estimator of μ. Assume that a second sample of size N-n is to be selected from the same distribution. Denote the mean of the second sample by $\bar{y}_{N-n}$. Then the best linear unbiased predictor of $\bar{y}_{N-n}$ is also $\bar{y}$. It follows that $\bar{y}_n$ is the best linear unbiased predictor of

$$\bar{Y} = N^{-1}[n\bar{y} + (N-n)\bar{y}_{N-n}] .$$

If we rearrange the timing of the sampling operations we can use this model as a framework for developing sample designs and estimators. Assume that we are first given the sample of N observations from the population. Then from the N we select a simple random sample of size n. The sample mean remains the best linear unbiased predictor of the mean of the remaining N-n elements. Considering the finite population to be a sample from a superpopulation is a way to formalize our prior information about the finite population. We can conceptualize the problem in this way at the design stage even if we plan to treat the finite population as a set of fixed numbers at the ultimate estimation stage.

At the design stage we assume that the finite population is a vector selected from an infinite population with mean $\underset{\sim}{\mu}$ and nonsingular NxN covariance matrix $\underset{\sim}{\Sigma}$, where $\underset{\sim}{\Sigma}$ has typical element σ_{ij}. In later formulations $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$ may be specified to be functions of unknown parameters.

We restrict designs to the class of random designs. In this formulation there are two sources of "variation" to be considered at the design stage. The first is associated with the superpopulation from which the finite population is viewed as a sample. The second is that introduced by the random sample design. We denote the expectation with respect to the superpopulation with the symbol $\mathcal{E}$ and the expectation with respect to the sampling design by E.

An estimator d is <u>conditionally model unbiased</u> for $\bar{Y}$ if

$$\mathcal{E}\{(d - \bar{Y}) | s\} = 0 . \tag{26}$$

The conditioning in (26) is with respect to the elements of s and the prior information for those elements, but not on the y-characteristics of s. The estimator d is said to be a <u>model unbiased predictor</u> of $\bar{Y}$ if (26) holds for all s in $\mathcal{S}$ such that $p\{s\} > 0$.

An estimator d is <u>unbiased for $\bar{Y}$ with respect to the design</u> p if

$$E\{d\} = \bar{Y} , \tag{27}$$

for all $(y_1, y_2, \ldots, y_N)$ contained in N-dimensional Euclidean space, where

$$E\{d\} = \Sigma_s d(s)\, p(s)$$