STATISTICS: textbooks and monographs volume 37

SAMPLING FROM A FINITE POPULATION

JAROSLAV HÁJEK

SAMPLING FROM A FINITE POPULATION

JAROSLAV HÁJEK
Charles University
Prague, Czechoslovakia

Edited by Václav Dupač
Charles University
Prague, Czechoslovakia

MARCEL DEKKER, INC. New York and Basel

Library of Congress Cataloging in Publication Data

Hajek, Jaroslav, [date]
Sampling from a finite population.

(Statistics, textbooks and monographs ; v. 37)
Bibliography: p.
Includes index.
1. Sampling (Statistics) I. Dupac, Vaclav.
II. Title. III. Series.
QA276.6.H334 519.5 81-7835
ISBN 0-8247-1291-9 AACR2

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

SAMPLING FROM A FINITE POPULATION

Jaroslav Hájek (1926-1974)

FOREWORD

The present monograph by the late professor Jaroslav Hájek (1926-1974) deals primarily with theory of sampling from a finite population, which plays an important and increasing role in statistical theory as well as in applications. In the majority of practical problems, finite population sampling underlies the statistical methodology, and the study of the various characteristics of the population in some desirable or optimal way, along with prescribed margins of error of the estimators, constitutes a basic domain of statistical planning and inference. The present monograph is devoted to these fundamental aspects of finite population sampling.

The late Professor Hájek earned the highest international reputation for his outstanding contributions to the theory of probability, parametric estimation, nonparametrics, statistical inference in stochastic processes, probabilistic sampling, as well as some other applied areas. Indeed, he was one of the most versatile statisticians, who not only contributed generously to various fields but also was able to provide natural bridges between them. In the sixties, he became deeply interested in the development of the theory of probabilistic sampling and made a valuable contribution to this area through a set of invaluable papers in the Annals of Mathematical Statistics and some other Czechoslovak journals. His last paper on sampling with varying probabilities without replacements in 1973 depicts his interest in this field until his untimely death in 1974.

In the early seventies, during his visit to the United States, he became interested in writing a monograph on finite population sampling and pursued the project almost to completion, in spite of his serious kidney problems. He literally worked on this project and guided research until the last days of his life. After his premature death, his colleagues and former students at the Charles University, Prague, Czechoslovakia, put together the final revisions of this monograph.

The publication of this monograph naturally was planned a few years ago, but was delayed, not only because of Professor Hájek's death, but also because of other factors, though the ideas contained in it continued to stimulate interest not only among colleagues at Prague but also in the United States as well as other parts of Europe. Mrs. Betty Hájková's sincere effort in putting together the last work of her late husband as a monograph has met with a good response (and profound respect).

I have no doubt in my mind that the monograph will be very useful to a variety of statisticians desiring to understand the basic principles of finite population sampling and to apply them fruitfully in practical situations. Like the other books of Hájek, the current one reveals his lucid style and presentation at an intermediate level of mathematical sophistication. I also believe that this monograph will stimulate further research in this area of great theoretical and practical interest .

P. K. Sen
Chapel Hill, North Carolina

CONTENTS

FOREWORD	P. K. Sen	iii
Part I	BASIC IDEAS AND CONCEPTS	1
	1. Population	3
	2. Bayesian Approach: Probability Speculation	14
	3. Robust Approach: Probability Sampling	20
	4. More About Inference	33
Part II	METHODS OF SAMPLING	47
	5. Simple Random Sampling	49
	6. Poisson Sampling	54
	7. Rejective Sampling	66
	8. Sampford-Durbin Modification of Rejective Sampling	85
	9. Successive Sampling	93
	10. Systematic Sampling	113
	11. Stratified Sampling	118
	12. Multistage Sampling	122
	13. Assorted Methods	126
	14. Conditional Poisson Sampling	132
	15. Correcting a Sample	144
Part III	METHODS OF ESTIMATION	151
	16. Representativeness, Robustness, and Tightness	153
	17. Simple Linear Estimate	162
	18. Optimum Sampling-Estimating Strategies	178
	19. Regression and Ratio Estimates	190
	20. Stratification	208
	21. Subsampling	221
	22. A General Approach	230
	23. Applications to Simple Random Sampling	236
REFERENCES		242
INDEX		245

Part I

BASIC IDEAS AND CONCEPTS

1

POPULATION

A collection of certain units is called a population if the units are not interesting by themselves but only as contributors to statistical properties of the whole. Thus, regarding an aggregate as a population, we do not ask which units have this or that attribute but only how many of them possess it.

Example 1.1 A book is not a mere population of words if we investigate its esthetic or informational contents. However, it becomes a population of words if we explore the frequency of pronouns or some other statistical property. Similarly a collection of workers in a factory is a population if the average age is considered, but is something more if the payroll is compiled.

Usually one is told that the units making up a population must be to a considerable degree similar. The positive meaning of this requirement is that there should be variables (characteristics) of interest that are applicable to all units making up the given population.

Given a population S and a variable y, the most trivial statistical issue is to establish the total

$$Y = y_1 + \cdots + y_N \tag{1.1}$$

where y_i denotes the value of y possessed by the ith unit. However, if the information about the values y_i is not easily accessible and if the number of units N is very large, the practical completion of this task may be rather expensive and time-absorbing. Then we may

exploit the fact that Y is a statistical characteristic and estimate it from a sample. This book will be devoted to discussion of various sampling and estimating strategies.

The units making up the population S may be any elements worth studying--persons, families, farms, account items, temperature readings, and so on--and their nature will be irrelevant for theoretical considerations. We shall assume that the units are identifiable by certain labels (tags, names, addresses) and that we have available a frame (list, map) showing how to reach any unit given its label. The simplest labels are the numbers 1, 2, ..., N, and as a matter of fact we shall identify the units with their ordinal numbers by putting $S = \{1,2,\ldots,N\}$.

Also the variable y may denote anything of interest: income, voting preference, number of children, temperature, and so on. In particular, y may indicate the presence or absence of a certain attribute A, that is, $y_i = 1$ if the unit i has the attribute, and $y_i = 0$ if it does not. In that case y will be called the *indicator* of the attribute A. Then, obviously

$$Y = y_1 + \cdots + y_N = N_A \tag{1.2}$$

where N_A is the number of units possessing the property A. We shall call N_A the *frequency* of A. A characteristic that classifies the units according to k mutually exclusive attributes (k different religions, for example) may be expressed by k corresponding indicators.

Example 1.2 In a population consisting of 180 men the circumference of their heads was measured. The results are presented in Table 1.1, where we can read that the total equals

$Y = 10{,}179$ cm

Example 1.3 Table 1.2 shows frequencies of 14 attributes characterizing dental state and health in the left lower quadrant of the mouth of 786 twelve-year-old girls in Czechoslovakia in the year 1955.

TABLE 1.1 Circumference of head in men

Circumference of head (cm)	Frequency	Product
52	1	52
53	3	159
54	11	594
55	22	1,210
56	56	3,136
57	35	1,995
58	41	2,378
59	6	354
60	4	240
61	1	61
Totals:	180	10,179

Dividing the total Y by the size N of the population, we obtain the *average*

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N}(y_1 + \cdots + y_N) \tag{1.3}$$

If y is the indicator of an attribute, then the average is identical with the *relative frequency* of the attribute in the population. We shall denote the relative frequency by P:

$$P = \frac{N_A}{N} \tag{1.4}$$

TABLE 1.2 Dental state and health in 12-year-old girls (N = 786)

Tooth number	1	2	3	4	5	6	7
Frequency of girls with the tooth cut through	782	785	779	759	686	785	702
Frequency of girls with the tooth cut through and healthy	760	770	777	745	617	109	379

Example 1.4 The average circumference of the head in men computed from the data of Table 1.1 equals

$$\bar{Y} = \frac{10,179}{180} = 56.6 \text{ cm}$$

The relative frequency of girls that have the tooth number 5 in the left lower quadrant cut through is, according to Table 1.2,

$$P = \frac{686}{786} = 0.873 = 87.3\%$$

The average is a special case of a ratio of two totals:

$$R = \frac{Y}{X} \tag{1.5}$$

where $X = x_1 + \cdots + x_N$ for some another variable x. Ratio characteristics are capable of expressing indices of growth, intensities, proportions--generally, the relative magnitude of y in terms of x.

Example 1.5 An important anthropometric characteristic is the ratio of the width and length of the face. The data given in Table 1.3 provide the following value for it:

$$R = \frac{\text{total of face lengths}}{\text{total of face widths}} = \frac{25,630}{21,936} = 1.17$$

Example 1.6 A natural way to judge the attribute "the tooth is cut through and healthy" is in terms of the attribute "the tooth is cut through." Table 1.4 provides values for the corresponding ratio characteristics for teeth numbers 1, 2, ..., 7.

The variability of y is usually measured by the *variance*

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2 \tag{1.6}$$

and, for nonnegative variables, the relative variability by the coefficient of variation

$$V_y = \frac{\sigma_y}{\bar{Y}} \tag{1.7}$$

TABLE 1.3 Men classified according to the width and length of the face (N = 179)

	Length (mm)											Conditional totals	
Width (mm)	131	134	137	140	143	146	149	152	155	158	161	Frequency	Length
108		1	1									2	271
111			1	1	1		1					4	569
114		1		2	2	2	2					9	1,290
117		3	4	9	9	9	3				1	38	5,419
120		1	6	4	9	6	2					28	3,977
123	1	1	3	6	11	3	3	1				29	4,126
126			2	3	3	6	10	2			1	27	3,954
129		2	6	6	8	2	4		1	1		30	4,275
132				1	3	1	1	1		1		8	1,174
135						1	1					2	295
138							1					1	149
141	1											1	131
Conditional totals:												Grand total:	
Frequency:	2	9	23	32	46	30	28	4	1	2	2	179	25,630
Width :	264	1,074	2,802	3,894	5,631	3,651	3,480	507	129	261	243	21,936	

TABLE 1.4 Percentage of healthy teeth among cut-through ones for 12-year-old girls and teeth Nos. 1, 2, ..., 7

Tooth number	Percentage
1	760/782 = 97.2%
2	770/785 = 98.1
3	777/779 = 99.7
4	745/759 = 98.2
5	617/686 = 89.9
6	109/785 = 13.9
7	379/702 = 54.0

Note: A tooth is qualified as healthy if free of cavities as well as of fillings.

The square root of σ_y^2, i.e., σ_y, is called the *standard deviation.* The joint variability of y and x may be expressed by the *covariance*

$$\sigma_{yx} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X}) \tag{1.8}$$

Dividing σ_{yx} by $\sigma_y \sigma_x$, we obtain the *correlation coefficient*

$$\rho = \frac{\sigma_{yx}}{\sigma_y \sigma_x} \tag{1.9}$$

which measures the extent of linear dependence of y on x.

Dividing by σ_x^2, we obtain the *coefficient of regression* of y on x:

$$\beta_{yx} = \frac{\sigma_{yx}}{\sigma_x^2} \tag{1.10}$$

We can see that the *regression line*

$$y = \bar{Y} + \beta_{yx}(x - \bar{X}) \tag{1.11}$$

passes through the origin, i.e., is equivalent to $y = \beta_{yx}x$, iff

$$\beta_{yx} = \frac{Y}{X} \tag{1.12}$$

If y is the indicator of an attribute A and $Y = N_A$, then the variance and the variation coefficient equal

$$\sigma_A^2 = \frac{N_A(N - N_A)}{N \cdot N} = P_A(1 - P_A) \tag{1.13}$$

$$V_A = \sqrt{\frac{N(N - N_A)}{N_A N}} = \sqrt{\frac{1}{P_A} - 1} \tag{1.14}$$

where $P_A = N_A/N$.

If y and x indicate attributes A and B, respectively, then their correlation coefficient equals

$$\rho_{AB} = \frac{P_{AB} - P_A P_B}{\sqrt{P_A(1 - P_A)P_B(1 - P_B)}} \tag{1.15}$$

where P_A, P_B, P_{AB} are relative frequencies of attributes A, B and of joint occurrence of A and B.

PROBLEMS AND NOTES

1.1 Compute the regression coefficient β_{yx} from Table 1.3 if y denotes the length and x denotes the width of the face. Show that it greatly differs from R = Y/X. Draw the regression line of y on x and observe that it passes far from the origin.

1.2 Compute β_{yx} and R = Y/X from the data of Table 1.5. Note that the regression nearly passes through the origin.

1.3 Let y and x indicate two attributes A and B such that A occurs only on those units where B does. Then the regression of y on x passes through the origin if $N_A > 0$.

1.4 Let y and x indicate two attributes that exclude each other. Then the correlation coefficient of y and x is negative. When does it equal -1?